Libri viventi: codicologia quantitativa, filometria dei manoscritti e algoritmi di stemmatica computazionale

Stemmatica computazionale e modelli evolutivi della trasmissione testuale

1.1 I manoscritti come organismi evolutivi: dalla metafora biologica alla modellazione quantitativa

La tradizione manoscritta dei testi antichi e medievali rappresenta uno dei fenomeni più affascinanti per l'applicazione di metodologie quantitative e computazionali alla ricerca umanistica. Ogni manoscritto che ci è pervenuto non è semplicemente un artefatto materiale che conserva un testo, ma costituisce piuttosto un anello di una catena genealogica complessa, il risultato di innumerevoli atti di copiatura, correzione, emendamento e trasmissione che si sono susseguiti nel corso dei secoli. Questa prospettiva, che considera i manoscritti come entità soggette a processi di trasmissione e mutazione analogamente a quanto avviene per gli organismi biologici, ha rivoluzionato la critica testuale trasformandola da disciplina prevalentemente qualitativa a scienza che integra rigorosi metodi quantitativi.

La metafora biologica applicata alla trasmissione testuale non è recente. Già nel diciottesimo e diciannovesimo secolo, i filologi classici avevano intuito che le relazioni tra i diversi testimoni manoscritti di un'opera potevano essere rappresentate mediante alberi genealogici, chiamati tecnicamente "stemma codicum". L'idea fondamentale è che ogni copia di un manoscritto deriva da uno o più esemplari precedenti, ereditandone non solo il testo corretto ma anche gli errori accumulatisi nel corso delle generazioni precedenti di copiatura. Questi errori, lungi dall'essere semplici disturbi da eliminare, costituiscono in realtà la chiave per ricostruire le relazioni genealogiche tra i manoscritti, esattamente come le mutazioni genetiche permettono di tracciare l'albero filogenetico delle specie viventi.

Karl Lachmann, nel diciannovesimo secolo, fu tra i primi a sistematizzare questo approccio, elaborando un metodo che prevedeva l'identificazione degli errori caratteristici di ciascun testimone manoscritto e la loro organizzazione in uno schema genealogico. Il metodo lachmanniano, tuttavia, era ancora largamente intuitivo e basato sul giudizio soggettivo del filologo. Solo con l'avvento della biologia evolutiva moderna e, successivamente, dell'informatica, è stato possibile sviluppare metodi rigorosamente quantitativi per la ricostruzione degli stemmi.

1.2 Lo stemma codicum: teoria e costruzione degli alberi genealogici dei manoscritti

Lo stemma codicum rappresenta la visualizzazione grafica delle relazioni genealogiche che intercorrono tra i diversi testimoni manoscritti di un'opera. In termini formali, uno stemma è un albero radicato in cui ciascun nodo rappresenta un manoscritto (o un ipotetico archetipo perduto) e ciascun arco rappresenta una relazione di discendenza diretta. Il nodo radice corrisponde all'archetipo, ossia il capostipite di tutta la tradizione manoscritta, mentre le foglie dell'albero corrispondono ai manoscritti effettivamente sopravvissuti fino ai nostri giorni.

La costruzione di uno stemma si basa sull'analisi sistematica delle varianti testuali, cioè delle differenze che si osservano confrontando i diversi testimoni. Non tutte le varianti, tuttavia, hanno lo stesso valore diagnostico per la ricostruzione dello stemma. La filologia tradizionale ha da tempo riconosciuto che solo gli errori significativi, chiamati "errori guida" o "errori separativi", possono essere utilizzati per stabilire le relazioni genealogiche. Un errore si definisce separativo quando divide la tradizione manoscritta in due gruppi: quello dei testimoni che lo presentano (e che quindi discendono da un antenato comune che aveva già introdotto quell'errore) e quello dei testimoni che non lo presentano.

La distinzione tra errori e innovazioni corrette è cruciale ma spesso problematica. In linea teorica, un errore è una lezione che si discosta dall'originale dell'autore, mentre una lezione corretta è quella che corrisponde al testo autentico. Nella pratica, tuttavia, non abbiamo accesso diretto all'originale (se lo avessimo, non avremmo bisogno di ricostruire lo stemma), e quindi dobbiamo basarci su criteri indiretti per distinguere le lezioni originali dalle corrotte. I criteri tradizionali includono la "lectio difficilior", il principio secondo cui, in caso di dubbio, la lezione più difficile o inusuale è da preferire, poiché i copisti tendevano a semplificare e banalizzare piuttosto che a complicare il testo.

Un concetto fondamentale nella teoria dello stemma è quello di "archetipo". L'archetipo non è necessariamente l'originale dell'autore, ma piuttosto il più antico antenato comune di tutti i manoscritti superstiti. Può darsi che tra l'originale e l'archetipo si siano interposte una o più generazioni di copie andate perdute. L'obiettivo della recensio (la fase di raccolta e classificazione dei testimoni) è proprio quello di ricostruire, per quanto possibile, il testo dell'archetipo, che rappresenta il punto più vicino all'originale a cui possiamo risalire sulla base della tradizione manoscritta superstite.

1.3 Cladistica e metodi filogenetici: l'applicazione dei principi di Hennig alla critica testuale

La rivoluzione metodologica nella stemmatica è avvenuta con l'importazione dei metodi della cladistica, la disciplina biologica sviluppata da Willi Hennig negli anni cinquanta del ventesimo secolo per ricostruire le relazioni filogenetiche tra le specie. Hennig propose un metodo rigorosamente sistematico basato sull'analisi dei caratteri condivisi derivati (sinapomorfie), che permetteva di distinguere tra raggruppamenti naturali (cladi) e raggruppamenti artificiali.

L'applicazione della cladistica alla critica testuale si basa sull'osservazione che la trasmissione dei testi segue principi analoghi a quelli della trasmissione biologica: ogni copia eredita le caratteristiche dell'esemplare da cui deriva, inclusi gli errori già presenti, e può introdurre nuovi errori che saranno poi trasmessi alle copie successive. Questa analogia è

resa ancora più stretta dal fatto che, in entrambi i casi, la trasmissione è prevalentemente verticale (da genitore a figlio, da antigrafo a copia) e le mutazioni sono irreversibili nella grande maggioranza dei casi.

Il primo passo nell'applicazione dei metodi cladistici consiste nella costruzione di una "character matrix" (matrice dei caratteri), una tabella in cui le righe rappresentano i diversi manoscritti e le colonne rappresentano i loci testuali in cui si osservano varianti. Ogni cella della matrice contiene lo stato del carattere per quel particolare manoscritto in quel particolare locus. Gli stati possono essere binari (0/1, dove tipicamente 0 indica la lezione originale e 1 l'innovazione) oppure multistatali quando in un dato locus si osservano tre o più varianti alternative.

La codifica delle varianti testuali in una matrice di caratteri richiede scelte metodologiche importanti. Innanzitutto, è necessario decidere quali varianti includere nell'analisi. Come accennato, non tutte le varianti sono ugualmente informative: le varianti puramente ortografiche o grafiche, per esempio, hanno scarso valore genealogico. In secondo luogo, occorre stabilire la polarità dei caratteri, ossia determinare quale stato rappresenti la condizione ancestrale e quale quella derivata. Questa operazione, chiamata "rooting" (radicamento) in cladistica, è fondamentale ma può presentare difficoltà, specialmente quando non disponiamo di un outgroup (un testimone esterno che sicuramente preservi la lezione originale).

Una volta costruita la character matrix, è possibile applicare vari algoritmi per inferire l'albero filogenetico più probabile. I metodi principali si dividono in tre categorie: metodi di parsimonia, metodi di massima verosimiglianza e metodi bayesiani.

1.4 Parsimonia e algoritmo di Wagner: minimizzare le mutazioni nello stemma

Il principio di parsimonia, o rasoio di Occam applicato alla filogenesi, afferma che l'albero migliore è quello che richiede il minor numero di mutazioni (cambiamenti evolutivi) per spiegare i dati osservati. In termini formali, dato un insieme di manoscritti e una character matrix, cerchiamo l'albero che minimizza la lunghezza totale, definita come la somma dei cambiamenti di stato richiesti su tutti i caratteri e su tutti i rami dell'albero.

L'algoritmo di Wagner è uno dei metodi più classici per trovare alberi parsimoniosi. L'algoritmo procede per addizione progressiva dei taxa (manoscritti): si parte da un albero contenente solo tre manoscritti, poi si aggiungono iterativamente i rimanenti, inserendo ciascun nuovo manoscritto nella posizione che minimizza l'incremento di lunghezza dell'albero. Per ogni possibile posizione di inserimento, si calcola la lunghezza dell'albero risultante utilizzando l'algoritmo di Fitch, che determina in modo ottimale gli stati dei nodi interni date le foglie e la topologia dell'albero.

Formalmente, la lunghezza L di un albero T per un singolo carattere si calcola come la somma minima dei cambiamenti di stato su tutti i rami. Se indichiamo con s(v) lo stato del carattere al nodo v, la lunghezza per quel carattere è:

$$L = \Sigma\, \delta(s(u), s(v))$$

dove la somma è estesa a tutti i rami (u,v) dell'albero, e δ è la distanza tra gli stati (δ = 0 se s(u) = s(v), δ = 1 se s(u) ≠ s(v) per caratteri binari). La lunghezza totale dell'albero è poi la somma delle lunghezze calcolate per tutti i caratteri nella matrice.

Il problema della ricerca dell'albero più parsimonioso è computazionalmente complesso (NP-hard), il che significa che per un numero elevato di taxa non è possibile esplorare esaustivamente tutte le possibili topologie. Per k taxa esistono infatti (2k-3)!! = (2k-3)(2k-5)...3×1 alberi radicati distinti, un numero che cresce in modo super-esponenziale. Per esempio, con 10 taxa ci sono circa 34 milioni di alberi possibili, con 20 taxa il numero supera 8×10^21. Per questa ragione si ricorre a euristiche di ricerca, come l'algoritmo di Wagner seguito da ottimizzazioni locali (branch swapping, subtree pruning and regrafting).

Un aspetto critico dell'analisi di parsimonia è la gestione dei caratteri omoplastici, ossia quei caratteri che mostrano lo stesso stato in taxa non strettamente imparentati a causa di evoluzioni parallele o convergenti, oppure a causa di reversioni. Nei manoscritti, l'omoplasia può derivare da errori poligenetici (lo stesso errore che si verifica indipendentemente in rami diversi dello stemma) oppure da contaminazione (vedi sezione successiva). Gli indici di consistenza (CI = m/s, dove m è il numero minimo di passi teoricamente necessario e s è il numero di passi osservato sull'albero) misurano il grado di omoplasia: un CI = 1 indica assenza di omoplasia, mentre valori inferiori indicano presenza di evoluzioni parallele o reversioni.

1.5 Maximum likelihood: modelli probabilistici di sostituzione testuale

I metodi di massima verosimiglianza (Maximum Likelihood, ML) offrono un approccio più sofisticato alla ricostruzione filogenetica, basato su modelli probabilistici espliciti del processo evolutivo. Invece di limitarsi a contare il numero minimo di mutazioni, i metodi ML calcolano la probabilità che i dati osservati (la character matrix) siano stati generati da un particolare albero, assumendo un modello specifico di evoluzione dei caratteri.

Il concetto fondamentale è la funzione di verosimiglianza L(T,θ|D), che rappresenta la probabilità dei dati D (la character matrix) dato un albero T con parametri θ. I parametri tipicamente includono le lunghezze dei rami (che rappresentano il "tempo evolutivo" o il numero atteso di mutazioni) e i parametri del modello di sostituzione (per esempio, il tasso di mutazione).

Per un singolo carattere e assumendo indipendenza tra siti, la verosimiglianza si calcola mediante l'algoritmo di Felsenstein, che sfrutta la struttura ad albero per computare efficientemente la probabilità mediante programmazione dinamica. Per ogni nodo interno v dell'albero, si definisce la probabilità parziale L_v(i) come la probabilità di osservare i dati in tutti i discendenti del nodo v, dato che il nodo v si trova nello stato i. Queste probabilità parziali si calcolano ricorsivamente dalle foglie verso la radice:

Per una foglia con stato osservato j: L_foglia(i) = 1 se i = j, 0 altrimenti

Per un nodo interno con figli u e w: L_v(i) = [Σ_j P_u(i,j) × L_u(j)] × [Σ_k P_w(i,k) × L_w(k)]

dove P_u(i,j) è la probabilità di transizione dallo stato i allo stato j lungo il ramo che connette v a u, calcolata secondo il modello di sostituzione adottato.

I modelli di sostituzione per i testi sono analoghi ai modelli usati in biologia molecolare, ma devono tener conto delle peculiarità della trasmissione manoscritta. Il modello più semplice è il modello di Jukes-Cantor generalizzato, che assume che tutte le sostituzioni abbiano la stessa probabilità. Per caratteri binari, la probabilità di transizione lungo un ramo di lunghezza t è:

$P(0\rightarrow0,t) = P(1\rightarrow1,t) = 1/2 + 1/2 \times \exp(-2\mu t)$ $P(0\rightarrow1,t) = P(1\rightarrow0,t) = 1/2 - 1/2 \times \exp(-2\mu t)$

dove μ è il tasso di mutazione. Modelli più complessi permettono di specificare tassi diversi per transizioni diverse, per esempio distinguendo tra innovazioni e reversioni, oppure tra diversi tipi di errori testuali (omissioni, sostituzioni, trasposizioni).

Una volta calcolata la verosimiglianza per tutti i caratteri (assumendo indipendenza tra siti, la verosimiglianza totale è il prodotto delle verosimiglianze individuali, o meglio la somma dei loro logaritmi per evitare underflow numerico), l'obiettivo è trovare l'albero e i parametri che massimizzano questa quantità. Anche in questo caso il problema è computazionalmente complesso, e si ricorre a euristiche di ricerca combinate con ottimizzazione numerica dei parametri (tipicamente mediante algoritmi come Newton-Raphson o Expectation-Maximization).

I metodi ML hanno il vantaggio di essere statisticamente ben fondati e di permettere il confronto formale tra modelli alternativi mediante test del rapporto di verosimiglianza (likelihood ratio test). Inoltre, forniscono misure di incertezza sulle stime (mediante l'analisi del profilo di verosimiglianza o bootstrap parametrico). Lo svantaggio principale è il costo computazionale elevato e la necessità di specificare un modello esplicito, che potrebbe non catturare adeguatamente le complessità della trasmissione testuale.

1.6 Filogenesi bayesiana: distribuzioni a posteriori e inferenza probabilistica

L'approccio bayesiano alla filogenesi rappresenta l'evoluzione più recente e sofisticata dei metodi statistici per la ricostruzione degli stemmi. A differenza dei metodi ML, che cercano il singolo albero "migliore" (quello che massimizza la verosimiglianza), i metodi bayesiani esplorano l'intero spazio delle possibili topologie e parametri, fornendo una distribuzione di probabilità a posteriori su questo spazio.

Il teorema di Bayes fornisce il fondamento teorico:

$P(T,\theta|D) = [P(D|T,\theta) \times P(T,\theta)] / P(D)$

dove $P(T,\theta|D)$ è la distribuzione a posteriori (ciò che vogliamo conoscere: la probabilità dell'albero T e dei parametri θ dati i dati), $P(D|T,\theta)$ è la verosimiglianza (la probabilità dei dati dato l'albero e i parametri), $P(T,\theta)$ è la distribuzione a priori (le nostre credenze preliminari su alberi e parametri), e $P(D)$ è l'evidenza o probabilità marginale dei dati (una costante di normalizzazione).

Il calcolo della distribuzione a posteriori non è analitico per problemi di questa complessità, ma può essere approssimato mediante metodi di Monte Carlo a catena di Markov (MCMC). L'idea è di costruire una catena di Markov il cui stato stazionario è la distribuzione a

posteriori desiderata, poi campionare da questa catena per ottenere un insieme rappresentativo di alberi e parametri.

L'algoritmo più utilizzato in filogenesi bayesiana è il Metropolis-Hastings accoppiato con mosse specifiche per esplorare lo spazio delle topologie. A ogni iterazione, si propone una modifica all'albero corrente (per esempio, un riarrangiamento della topologia o una modifica delle lunghezze dei rami), si calcola il rapporto di accettazione:

$$\alpha = \min\{1, [P(D|T',\theta') \times P(T',\theta') \times q(T,\theta|T',\theta')] / [P(D|T,\theta) \times P(T,\theta) \times q(T',\theta'|T,\theta)]\}$$

dove T,θ rappresentano lo stato corrente e T',θ' lo stato proposto, e q è la distribuzione di proposta. Si accetta la mossa con probabilità α, altrimenti si mantiene lo stato corrente. Dopo un periodo di burn-in (le prime iterazioni, che permettono alla catena di raggiungere la distribuzione stazionaria), i campioni successivi approssimano la distribuzione a posteriori.

Il software BEAST (Bayesian Evolutionary Analysis Sampling Trees) è uno degli strumenti più utilizzati per l'analisi filogenetica bayesiana. BEAST permette di specificare modelli evolutivi complessi, inclusi modelli con tassi di mutazione variabili nel tempo (relaxed clock models) e modelli di speciazione/estinzione. Nel contesto della stemmatica, BEAST può essere utilizzato per integrare informazioni cronologiche (per esempio, le date dei manoscritti basate su evidenze paleografiche o storiche) mediante la calibrazione dell'orologio molecolare.

Un vantaggio fondamentale dell'approccio bayesiano è che fornisce misure naturali di incertezza su tutti gli aspetti dell'inferenza. Per esempio, possiamo calcolare le probabilità a posteriori dei cladi (gruppi monofiletici), che indicano il supporto statistico per ciascun raggruppamento di manoscritti. Alberi con supporto debole per alcuni cladi suggeriscono che i dati sono insufficienti o contraddittori riguardo a quelle relazioni genealogiche, segnalando la necessità di cautela nell'interpretazione o di ulteriori indagini.

1.7 Contaminazione e trasmissione orizzontale: quando gli alberi non bastano

Una delle assunzioni fondamentali dei metodi filogenetici tradizionali è che la trasmissione sia strettamente verticale: ogni copia deriva da un singolo antigrafo, ereditandone tutte le caratteristiche. Nella realtà della trasmissione manoscritta, tuttavia, questa assunzione è spesso violata. Il fenomeno della contaminazione si verifica quando un copista utilizza due o più esemplari come fonte, combinandone le lezioni, oppure quando corregge la propria copia consultando un altro manoscritto appartenente a un ramo diverso dello stemma.

La contaminazione rappresenta una forma di trasmissione orizzontale, analoga al trasferimento genico orizzontale in biologia (dove organismi non direttamente imparentati scambiano materiale genetico). Questo fenomeno introduce reticolazioni nello stemma: invece di un semplice albero, otteniamo un network in cui alcuni nodi hanno più di un genitore.

La presenza di contaminazione si manifesta attraverso segnali conflittuali nei dati: diversi caratteri suggeriscono alberi incompatibili tra loro. Per esempio, un manoscritto M potrebbe condividere alcuni errori con il ramo A dello stemma e altri errori con il ramo B, suggerendo

che il copista ha utilizzato fonti da entrambi i rami. Questi segnali conflittuali sono problematici per i metodi basati su alberi, che cercano di forzare i dati in una topologia strettamente dicotomica.

I metodi di network filogenetici affrontano esplicitamente questa complessità. Invece di cercare un singolo albero ottimale, questi metodi costruiscono reti che possono rappresentare relazioni reticolate. L'algoritmo NeighborNet, per esempio, costruisce un "splits graph" che visualizza tutti i raggruppamenti (splits) supportati dai dati. In un splits graph, i segnali conflittuali appaiono come parallelogrammi o box nella rete: quanto più ampio è il box, tanto più forte è il conflitto tra i segnali.

Formalmente, uno split è una bipartizione dell'insieme dei taxa in due sottoinsiemi complementari. Per esempio, con i manoscritti A, B, C, D, lo split AB|CD divide i taxa in {A,B} e {C,D}. Uno split è compatibile con un albero se esiste un ramo nell'albero la cui rimozione produce esattamente quella bipartizione. In un albero, tutti gli split devono essere mutuamente compatibili; in una rete, possono coesistere split incompatibili.

L'algoritmo NeighborNet funziona come segue: si parte da una matrice di distanze tra tutti i manoscritti (calcolate per esempio usando la distanza di Hamming sulla character matrix: il numero di siti in cui due manoscritti differiscono). Poi si identificano iterativamente le "vicinanze" (neighbors) nella rete, ossia i gruppi di taxa che sembrano più strettamente correlati secondo le distanze. A differenza del metodo Neighbor-Joining, che produce un albero, NeighborNet mantiene tutte le vicinanze, anche quelle conflittuali, rappresentandole come splits nel grafo finale.

Il risultato è una visualizzazione che mostra sia le relazioni ad albero (rappresentate da rami dicotomici) sia le regioni di conflitto (rappresentate da reticolazioni). L'interpretazione di questi grafici richiede competenza, ma essi forniscono informazioni preziose sulla struttura della tradizione manoscritta, evidenziando dove la contaminazione o l'evoluzione parallela ha creato segnali conflittuali.

Un approccio alternativo è quello dei modelli espliciti di trasmissione mista. Questi modelli permettono di stimare simultaneamente la topologia dell'albero principale e i parametri di contaminazione (quali manoscritti hanno attinto da fonti multiple, e in quale proporzione). Per esempio, si può modellare un manoscritto contaminato come una miscela: una percentuale α del testo deriva dal ramo A, una percentuale 1-α dal ramo B. Questi modelli sono computazionalmente molto impegnativi ma rappresentano la frontiera della ricerca in stemmatica quantitativa.

Critica testuale quantitativa e modellazione stocastica della trasmissione

2.1 Textual criticism: ricostruire l'archetipo attraverso l'analisi delle varianti

La critica testuale è la disciplina che si occupa di ricostruire, nella misura del possibile, il testo originale di un'opera a partire dai testimoni manoscritti superstiti, tutti inevitabilmente

corrotti in vari gradi. Il compito è reso difficile dal fatto che l'originale dell'autore è quasi sempre perduto, e dobbiamo basarci su copie di copie, ciascuna delle quali ha introdotto nuovi errori nel processo di trasmissione.

Il metodo critico tradizionale si articola in tre fasi principali: la recensio (raccolta e classificazione dei testimoni mediante la costruzione dello stemma), l'examinatio (esame delle varianti per determinare quale lezione sia originale in ciascun locus), e l'emendatio (correzione congetturale dei passi in cui tutti i testimoni sembrano corrotti). L'approccio quantitativo fornisce strumenti rigorosi soprattutto per le prime due fasi, mentre l'emendatio congetturale rimane in larga misura un'arte che richiede profonda conoscenza della lingua, dello stile dell'autore e del contesto culturale.

Una volta costruito lo stemma mediante i metodi discussi nella sezione precedente, la ricostruzione del testo dell'archetipo procede applicando sistematicamente il principio della maggioranza genealogica: quando i testimoni si dividono su una variante, si segue la lezione attestata dalla maggioranza dei rami indipendenti dello stemma, non semplicemente dalla maggioranza dei manoscritti. Questo perché ciò che conta è il numero di tradizioni indipendenti che supportano una lezione, non il numero di copie meccanicamente dipendenti da un singolo antenato.

Per illustrare questo punto con un esempio concreto, supponiamo di avere uno stemma in cui l'archetipo α ha generato due iparchetipi β e γ. Da β derivano i manoscritti A, B, C, mentre da γ derivano solo D ed E. Se in un certo locus i manoscritti A, B, C presentano la lezione "templum" mentre D ed E presentano "fanum", non possiamo concludere automaticamente che "templum" sia la lezione dell'archetipo α sulla base della maggioranza numerica (3 contro 2). Dobbiamo invece considerare che A, B, C costituiscono un'unica testimonianza (quella di β), mentre D ed E costituiscono un'altra testimonianza indipendente (quella di γ). In questo caso abbiamo una situazione di parità (1:1), e dobbiamo ricorrere ad altri criteri.

Il principio della lectio difficilior probabilior (la lezione più difficile è probabilmente quella originale) è il criterio più importante quando lo stemma non fornisce una risposta univoca. L'idea è che i copisti, nel corso della trasmissione, tendevano a semplificare, banalizzare, correggere ciò che appariva oscuro o grammaticalmente scorretto. Di conseguenza, quando ci troviamo di fronte a due varianti, quella più inusuale, arcaica, difficile dal punto di vista sintattico o lessicale, o apparentemente scorretta, ha maggiori probabilità di essere originale, poiché è più verosimile che un copista abbia trasformato il difficile in facile piuttosto che il contrario.

Tuttavia, questo principio deve essere applicato con cautela e competenza filologica. Non sempre la lezione più difficile è quella autentica: potrebbe essere il risultato di un errore che ha reso incomprensibile un passo originariamente chiaro. Inoltre, la valutazione di cosa costituisca una "lezione difficile" richiede profonda conoscenza dell'uso linguistico dell'epoca, del genere letterario, dello stile dell'autore. Una lezione che appare inusuale a un filologo moderno potrebbe essere stata perfettamente normale nel contesto dell'autore, e viceversa.

Altri criteri interni includono la lectio brevior (la lezione più breve è probabilmente originale, poiché i copisti tendevano ad aggiungere piuttosto che omettere), l'usus scribendi dell'autore (la lezione conforme alle abitudini stilistiche e lessicali dell'autore ha maggiori probabilità di

essere autentica), e la coerenza contestuale (la lezione che si armonizza meglio con il contesto immediato e l'argomentazione complessiva). L'applicazione di questi criteri richiede il "fiuto" del filologo esperto, ma può essere supportata da analisi quantitative dello stile e del vocabolario dell'autore.

2.2 Modelli stocastici della corruzione testuale: processi di nascita-morte

La trasmissione dei testi può essere modellata come un processo stocastico in cui errori vengono introdotti, accumulati e talvolta corretti nel corso delle generazioni successive di copiatura. I modelli stocastici forniscono un quadro matematico rigoroso per comprendere e quantificare questi fenomeni.

Il modello più semplice concettualizza la corruzione testuale come un processo di nascita puro: a ogni atto di copiatura, nuovi errori vengono introdotti (nascono) e si sommano agli errori preesistenti, mentre non si verificano correzioni (assenza di morte degli errori). Se indichiamo con N_t il numero di errori presenti dopo t generazioni di copiatura, e assumiamo che a ogni generazione vengano introdotti in media λ errori, il numero atteso di errori cresce linearmente con il tempo:

E[N_t] = λt

La distribuzione di probabilità di N_t segue una distribuzione di Poisson con parametro λt:

P(N_t = k) = (λt)^k × exp(-λt) / k!

Questo modello è evidentemente troppo semplificato per la maggior parte delle situazioni reali. Un'estensione naturale è il processo di nascita-morte, che tiene conto sia dell'introduzione di nuovi errori (nascita, con tasso λ) sia della correzione di errori preesistenti (morte, con tasso μ). In questo caso, la dinamica è governata dalle equazioni:

dE[N_t]/dt = λ - μN_t

dove λ rappresenta il tasso assoluto di introduzione di errori (errori per generazione) e μ il tasso relativo di correzione (frazione di errori corretti per generazione). La soluzione di questa equazione differenziale fornisce il numero atteso di errori al tempo t:

E[N_t] = (λ/μ) × [1 - exp(-μt)]

che tende asintoticamente a un valore di equilibrio λ/μ. Questo modello cattura l'osservazione empirica che, almeno in alcune tradizioni manoscritte caratterizzate da attività filologica e revisione critica, il tasso di corruzione non cresce indefinitamente ma tende a stabilizzarsi.

Un aspetto cruciale è la distinzione tra errori reversibili e irreversibili. Molti errori sono praticamente irreversibili: una volta che un passaggio è stato gravemente corrotto, è estremamente improbabile che un copista lo ricostruisca spontaneamente nella forma corretta. Altri errori sono più facilmente correggibili, specialmente se il copista dispone di

cultura e competenza per emendare ciò che gli appare manifestamente scorretto. I modelli più sofisticati distinguono tra classi di errori con diverse probabilità di reversione.

2.3 Tassi di errore e fedeltà di copiatura: evidenze empiriche

La quantificazione empirica dei tassi di errore nella trasmissione manoscritta è fondamentale per calibrare i modelli stocastici e per comprendere le dinamiche della corruzione testuale. Gli studi su questo tema hanno analizzato casi in cui disponiamo sia dell'esemplare utilizzato come antigrafo sia della copia che ne deriva, permettendo un confronto diretto e una misurazione precisa degli errori introdotti.

Definiamo il tasso di errore ε come il numero di errori per unità di testo (tipicamente per pagina o per 1000 parole) introdotti in un singolo atto di copiatura. Gli studi empirici hanno rivelato una notevole variabilità in questo parametro, in funzione di diversi fattori: la competenza e l'attenzione del copista, la leggibilità dell'antigrafo, la difficoltà del testo, il contesto istituzionale della copiatura.

Per gli scriptoria monastici del Medioevo, caratterizzati da copisti professionali che lavoravano in condizioni controllate e con una forte etica della fedeltà al testo, i tassi di errore sono relativamente bassi, tipicamente nell'intervallo ε ≈ 0,001 - 0,01 errori per pagina. Questo significa che un copista attento poteva trascrivere un'intera pagina commettendo in media solo un errore ogni 100-1000 pagine, un livello di accuratezza notevole considerando le difficoltà paleografiche e l'assenza di ausili moderni.

Al contrario, per i copisti commerciali del tardo Medioevo e del Rinascimento, che lavoravano spesso sotto pressione economica e con minore formazione, i tassi di errore sono significativamente più elevati, tipicamente ε ≈ 0,01 - 0,1 errori per pagina o anche superiori. In casi di copiatura frettolosa o negligente, si possono osservare tassi ancora più alti.

La propagazione degli errori attraverso le generazioni successive di copiatura può essere modellata come segue. Se ogni copia introduce in media ε errori per pagina, dopo n generazioni di copiatura la frazione di pagine ancora corrette (senza alcun errore) è approssimativamente:

P(pagina corretta dopo n generazioni) = (1-ε)^n ≈ exp(-nε)

dove l'approssimazione è valida per ε piccolo. Questo mostra che la corruzione del testo segue un andamento esponenziale: il testo si degrada progressivamente, e la probabilità che un passaggio rimanga intatto diminuisce esponenzialmente con il numero di generazioni.

Per fare un esempio concreto, consideriamo un manoscritto di 200 pagine copiato con un tasso di errore ε = 0,01. Dopo 10 generazioni di copiatura, la probabilità che una data pagina sia ancora completamente corretta è circa (1-0,01)^10 ≈ 0,904, ossia circa il 90%. Questo potrebbe sembrare rassicurante, ma significa anche che in media 20 pagine su 200 conterranno almeno un errore. Dopo 50 generazioni, questa probabilità scende a circa 0,605, e dopo 100 generazioni a 0,366. In altre parole, dopo un secolo di trasmissione

manoscritta (ipotizzando una generazione ogni anno, un'ipotesi plausibile per testi molto diffusi), più della metà delle pagine conterrà almeno un errore.

Questi calcoli evidenziano l'importanza della conservazione di testimoni antichi. Un manoscritto che si trova solo poche generazioni di distanza dall'archetipo è preziosissimo, poiché avrà accumulato molti meno errori rispetto a copie tarde. Questo principio è ben noto ai filologi sotto il nome di recentiores non deteriores (i testimoni più recenti non sono necessariamente peggiori), che tuttavia deve essere inteso correttamente: un testimone recente può essere prezioso se deriva da un antigrafo antico ora perduto, ma in generale la probabilità di corruzione aumenta con il numero di generazioni di trasmissione.

2.4 Fedeltà di copiatura: variabilità individuale e fattori contestuali

Il tasso di errore ε non è una costante universale ma varia considerevolmente in funzione di molteplici fattori. La competenza del copista è ovviamente cruciale: un copista esperto, con buona conoscenza della lingua e possibilmente del contenuto del testo, commetterà molti meno errori di un copista inesperto o che copia meccanicamente senza comprendere ciò che trascrive.

La leggibilità dell'antigrafo è un altro fattore determinante. Un manoscritto con scrittura chiara e regolare è molto più facile da copiare accuratamente rispetto a uno con scrittura corsiva, abbreviazioni estese, cancellature e correzioni. Gli studi paleografici hanno documentato come l'evoluzione delle scritture librarie verso forme sempre più formalizzate e standardizzate (come la carolina nel IX secolo o la littera textualis nel XIII secolo) abbia contribuito significativamente a ridurre i tassi di errore.

Il tipo di testo influenza anch'esso la fedeltà di copiatura. I testi sacri, come la Bibbia, erano copiati con particolare cura e reverenza, risultando in tassi di errore generalmente bassi. I testi liturgici, utilizzati quotidianamente nelle celebrazioni, dovevano essere accurati per garantire la correttezza del rito. Al contrario, opere letterarie o scientifiche potevano essere trattate con minor rigore, specialmente se destinate alla consultazione privata piuttosto che all'uso pubblico.

Il contesto istituzionale gioca un ruolo fondamentale. Gli scriptoria monastici, organizzati secondo regole precise che prescrivevano ritmi di lavoro sostenibili, controllo di qualità, correzione delle bozze, tendevano a produrre copie più accurate. I copisti lavoravano in ambienti silenziosi e ben illuminati, con pause regolari per evitare l'affaticamento. Spesso un monaco esperto rivedeva il lavoro del copista, segnalando gli errori da correggere.

Gli scriptoria commerciali del tardo Medioevo, invece, operavano sotto pressioni economiche: il copista era pagato a pagina o a quaderno, creando un incentivo a lavorare velocemente piuttosto che accuratamente. Questo sistema di produzione, pur avendo il merito di ampliare notevolmente la circolazione dei libri, ha comportato un generale declino nella qualità media delle copie.

Un aspetto interessante emerso dagli studi quantitativi è la presenza di effetti di apprendimento e affaticamento. Analizzando manoscritti lunghi, si osserva spesso che il tasso di errore è leggermente più alto nelle prime pagine (quando il copista si sta ancora familiarizzando con la scrittura dell'antigrafo e con lo stile del testo) e nelle ultime pagine

(quando subentra l'affaticamento). Il tasso di errore minimo si osserva tipicamente nella parte centrale del lavoro.

2.5 Propagazione degli errori e correzione: dinamiche complesse

La dinamica della corruzione testuale è ulteriormente complicata dal fatto che i copisti non solo introducevano nuovi errori, ma talvolta correggevano errori presenti nell'antigrafo. Queste correzioni potevano avvenire in vari modi: consultando un altro esemplare (contaminazione), basandosi sulla conoscenza del testo o della lingua (emendatio ope ingenii), o semplicemente per congettura fortunata.

Le correzioni complicate notevolmente l'analisi stemmatica perché introducono reversioni: un manoscritto che in un certo locus presenta la lezione corretta potrebbe averla non per averla ereditata direttamente dall'archetipo, ma per aver corretto un errore presente nell'antigrafo. Queste reversioni possono indurre in errore gli algoritmi filogenetici che assumono l'irreversibilità delle mutazioni.

La probabilità di correzione di un errore dipende da vari fattori. Gli errori che producono un testo manifestamente scorretto o privo di senso hanno maggiori probabilità di essere notati e corretti da un copista attento. Al contrario, gli errori che producono un testo apparentemente plausibile, anche se non originale, tendono a passare inosservati e a propagarsi nelle generazioni successive. Questi ultimi sono particolarmente insidiosi per la critica testuale.

Un tipo di errore particolarmente interessante dal punto di vista della modellazione è l'errore poligenetico: un errore che si verifica indipendentemente in più rami dello stemma. Certi tipi di errori sono statisticamente molto probabili: per esempio, la confusione tra lettere paleograficamente simili (come u e n nella minuscola carolina, o c e t nella minuscola gotica), l'aplografia (omissione di una lettera o sillaba doppia), la dittografia (duplicazione indebita di una lettera o sillaba). Questi errori possono manifestarsi indipendentemente in copisti diversi che affrontano lo stesso passaggio problematico.

La presenza di errori poligenetici complica la ricostruzione dello stemma perché viola l'assunzione di monogenesi degli errori: non è più vero che tutti i testimoni che condividono un certo errore derivano necessariamente da un antenato comune che aveva già introdotto quell'errore. I metodi statistici moderni cercano di tenere conto di questo fenomeno assegnando pesi diversi agli errori in funzione della loro probabilità di poligenesi.

2.6 Modelli gerarchici e variabilità tra loci

I modelli discussi finora assumono implicitamente che il tasso di errore sia uniforme attraverso tutto il testo. Nella realtà, tuttavia, esistono notevoli variazioni da un locus all'altro. Alcuni passaggi sono paleograficamente più difficili da leggere, altri sono linguisticamente più complessi, altri ancora contengono nomi propri o termini tecnici non familiari al copista.

I modelli gerarchici permettono di catturare questa eterogeneità. Invece di un singolo parametro ε globale, postuliamo una distribuzione di tassi di errore ε_i per i diversi loci i del testo. Per esempio, potremmo assumere che i tassi di errore seguano una distribuzione gamma con parametri di forma e scala da stimare dai dati:

$\varepsilon_i \sim \text{Gamma}(\alpha, \beta)$

Questo approccio è concettualmente analogo ai modelli "gamma di tassi" utilizzati in filogenesi molecolare per catturare la variazione nei tassi di sostituzione tra siti del DNA. I parametri α e β controllano rispettivamente la forma e la scala della distribuzione: valori bassi di α indicano forte eterogeneità (alcuni loci molto più propensi agli errori di altri), mentre valori alti indicano relativa uniformità.

L'inferenza in questi modelli gerarchici è più complessa ma fornisce stime più realistiche e robuste delle relazioni genealogiche. Inoltre, l'identificazione dei loci con alti tassi di errore può fornire informazioni utili sui processi di trasmissione e sulle difficoltà specifiche incontrate dai copisti.

Un'altra fonte di eterogeneità è la variazione tra copisti. Se un manoscritto è stato copiato da più copisti che si sono alternati (come spesso accadeva per opere molto lunghe), possiamo aspettarci tassi di errore diversi nelle sezioni redatte da copisti diversi. L'identificazione dei cambi di mano attraverso l'analisi paleografica permette di segmentare il manoscritto e di stimare tassi di errore specifici per ciascun copista, fornendo informazioni sulla variabilità individuale nella competenza scrittoria.

Paleografia quantitativa, datazione dei manoscritti e analisi del ductus scrittorio

3.1 Paleografia quantitativa: dall'analisi qualitativa alla misurazione oggettiva

La paleografia è la disciplina che studia le scritture antiche, i loro caratteri distintivi, la loro evoluzione nel tempo e nello spazio. Tradizionalmente, la paleografia è stata una scienza eminentemente qualitativa, basata sull'occhio esperto del paleografo che riconosce intuitivamente le caratteristiche distintive di una scrittura e le colloca nel tempo e nello spazio appropriati. Negli ultimi decenni, tuttavia, si è sviluppato un crescente interesse per l'applicazione di metodi quantitativi all'analisi paleografica, con l'obiettivo di rendere le valutazioni più oggettive, riproducibili e suscettibili di elaborazione statistica.

La paleografia quantitativa si basa sulla misurazione sistematica di parametri morfometrici della scrittura. Questi parametri includono caratteristiche geometriche come altezze, larghezze, angoli, curvature, proporzioni, spaziature. Per esempio, si può misurare l'altezza delle aste ascendenti rispetto all'altezza del corpo della lettera (il cosiddetto modulo), l'angolo di inclinazione delle aste rispetto alla linea di base, la larghezza media delle lettere, il rapporto tra pieni e filetti (le parti più spesse e più sottili del tratto di penna), la regolarità della spaziatura tra lettere e tra parole.

Il ductus è il movimento caratteristico della mano che produce una determinata forma di scrittura. Include aspetti come la sequenza dei tratti che compongono ciascuna lettera, la direzione del movimento (da sinistra a destra, dall'alto verso il basso), la velocità di esecuzione, la pressione esercitata sullo strumento scrittorio. Molti di questi aspetti possono

essere inferiti dall'esame del risultato finale sulla pagina: per esempio, l'angolo della penna rispetto alla linea di base può essere dedotto dall'orientamento dei tratti più spessi.

L'analisi quantitativa del ductus procede tipicamente come segue. Si seleziona un campione rappresentativo di lettere dal manoscritto, coprendo tutte le lettere dell'alfabeto e possibilmente diverse occorrenze di ciascuna lettera per catturare la variabilità intra-scrittore. Per ciascuna lettera campionata, si misurano vari parametri morfometrici utilizzando software di analisi di immagini. Questi parametri vengono poi sottoposti ad analisi statistica per identificare pattern caratteristici.

Un parametro particolarmente informativo è l'angolo della penna, che può essere stimato misurando l'orientamento dei tratti verticali o obliqui. Nelle scritture formali medievali, come la carolina o la gotica textualis, l'angolo della penna era relativamente standardizzato e costituiva un elemento distintivo della tipologia scrittoria. Per la carolina, l'angolo tipico è di circa 30-45 gradi rispetto all'orizzontale, mentre per la gotica textualis è più ripido, circa 40-50 gradi.

La misurazione dell'angolo della penna può essere automatizzata utilizzando algoritmi di analisi di immagini. Un approccio comune consiste nel binarizzare l'immagine della scrittura (convertendola in bianco e nero), identificare i tratti individuali mediante operazioni morfologiche (erosione, dilatazione, scheletrizzazione), e calcolare per ciascun tratto l'orientazione principale mediante analisi delle componenti principali o trasformata di Hough. La distribuzione degli angoli così ottenuti fornisce una caratterizzazione quantitativa del ductus.

Altri parametri morfometrici includono il rapporto di aspetto delle lettere (altezza/larghezza), che varia considerevolmente tra diverse tipologie scrittorie (la gotica textualis tende ad avere lettere compresse orizzontalmente, con rapporti di aspetto elevati, mentre la carolina ha forme più tondeggianti e rapporti più vicini a 1), la regolarità della spaziatura (misurabile come coefficiente di variazione delle distanze tra lettere successive), la densità di scrittura (numero di lettere per unità di area, correlato con l'economicità dell'uso della pergamena).

3.2 Machine learning per la classificazione paleografica

L'applicazione del machine learning alla paleografia ha aperto nuove possibilità per la classificazione automatica e la datazione dei manoscritti. Gli algoritmi di apprendimento automatico possono essere addestrati su insiemi di manoscritti datati con certezza (per esempio, manoscritti che riportano esplicitamente la data di copia nel colophon, o manoscritti databili con precisione attraverso evidenze storiche o codicologiche) per imparare a riconoscere i pattern morfometrici caratteristici di diversi periodi o regioni.

Il processo tipico di classificazione mediante machine learning si articola nelle seguenti fasi:

1. Preparazione del training set: si raccoglie un corpus di manoscritti datati con certezza, rappresentativo della varietà di tipologie scrittorie e periodi che si vogliono distinguere.

2. Estrazione delle features: per ciascun manoscritto del training set, si estraggono features quantitative che caratterizzano la scrittura. Queste possono includere i

parametri morfometrici discussi sopra, ma anche features più sofisticate derivate dall'analisi di texture, dalla trasformata wavelet, da descrittori di forma come i descrittori di Fourier.

3. Addestramento del modello: si utilizza un algoritmo di machine learning (per esempio, Support Vector Machines, Random Forests, reti neurali) per apprendere la relazione tra le features e le etichette (date o categorie paleografiche). Il modello impara a discriminare tra diverse classi sulla base dei pattern presenti nelle features.

4. Validazione e testing: le prestazioni del modello vengono valutate su un insieme di test indipendente (manoscritti non utilizzati durante l'addestramento) per verificare la capacità di generalizzazione.

5. Applicazione a manoscritti non datati: il modello addestrato può essere utilizzato per classificare o datare manoscritti di cui non si conosce la data di produzione.

Un approccio particolarmente promettente è l'uso di reti neurali convoluzionali (Convolutional Neural Networks, CNN), che hanno rivoluzionato il campo del riconoscimento di immagini. Le CNN possono apprendere automaticamente features rilevanti direttamente dalle immagini delle scritture, senza bisogno di specificare manualmente quali caratteristiche morfometriche misurare. Questo è particolarmente vantaggioso perché permette di catturare aspetti della scrittura che potrebbero sfuggire all'analisi tradizionale ma che sono discriminativi per la classificazione.

L'architettura tipica di una CNN per la classificazione paleografica consiste in una serie di strati convoluzionali che estraggono features gerarchiche (dai bordi e contorni nei primi strati, fino a pattern più complessi e astratti negli strati successivi), alternati con strati di pooling che riducono la dimensionalità, seguiti da strati completamente connessi che effettuano la classificazione finale. L'addestramento avviene mediante backpropagation, minimizzando una funzione di loss (tipicamente la cross-entropy per problemi di classificazione).

I risultati ottenuti con questi metodi sono incoraggianti. Studi recenti hanno dimostrato che le CNN possono raggiungere accuratezze del 80-90% nella classificazione di scritture in categorie paleografiche ampie (per esempio, carolina vs gotica), e accuratezze del 60-80% nella datazione entro intervalli di 25-50 anni. Queste prestazioni, pur non raggiungendo ancora quelle di un paleografo esperto, sono notevoli considerando la difficoltà intrinseca del compito e l'ampia variabilità anche tra manoscritti contemporanei.

3.3 Analisi della spaziatura e ritmo scrittorio

Un aspetto della scrittura che ha ricevuto crescente attenzione nell'ambito della paleografia quantitativa è la spaziatura, sia tra lettere all'interno delle parole sia tra parole. La spaziatura è un carattere distintivo di diverse tradizioni scrittorie e può fornire informazioni importanti per la datazione e la localizzazione.

Nelle scritture più antiche, come le maiuscole romane o le onciali, la separazione tra parole era spesso assente o inconsistente (scriptio continua). L'introduzione di una spaziatura regolare tra parole è un'innovazione che si afferma gradualmente nel corso del Medioevo, con tempi e modalità diverse nelle varie regioni europee. L'analisi quantitativa della spaziatura può quindi contribuire a tracciare questa evoluzione.

La spaziatura può essere quantificata misurando le distanze tra lettere successive e tra parole successive. Per ciascuna coppia di lettere consecutive, si misura la distanza minima tra i loro profili (la distanza più breve tra qualsiasi pixel appartenente alla prima lettera e qualsiasi pixel appartenente alla seconda). Si ottiene così una distribuzione di distanze inter-lettera. Analogamente, si misurano le distanze tra parole successive.

In una scrittura ben regolata, la distribuzione delle distanze inter-lettera dovrebbe essere relativamente stretta e centrata attorno a un valore medio caratteristico, mentre le distanze inter-parola dovrebbero formare una distribuzione separata, con media significativamente maggiore. La separazione tra queste due distribuzioni quantifica la chiarezza con cui le parole sono delimitate visualmente. Scritture con scarsa separazione (distribuzione bimodale poco pronunciata) corrispondono a una spaziatura poco sistematica, mentre scritture con netta separazione indicano una consapevole distinzione tra parole.

Il coefficiente di variazione della spaziatura (rapporto tra deviazione standard e media) è un indice della regolarità del ritmo scrittorio. Valori bassi indicano grande uniformità nella spaziatura, caratteristica delle scritture formali e calligrafiche prodotte da copisti esperti che mantenevano un ritmo costante. Valori elevati suggeriscono maggiore irregolarità, tipica di scritture corsive, rapide, o di copisti meno abili.

Analisi sofisticate possono anche esaminare se la spaziatura varia sistematicamente in funzione della posizione nella riga (per esempio, molti copisti tendevano a comprimere leggermente le lettere verso la fine della riga per evitare sillabazioni scomode) o in funzione di combinazioni specifiche di lettere (alcune coppie di lettere richiedono naturalmente più o meno spazio per evitare sovrapposizioni o vuoti eccessivi).

3.4 Datazione bayesiana dei manoscritti: integrare evidenze multiple

La datazione dei manoscritti è uno dei compiti fondamentali della codicologia, ma è anche uno dei più difficili quando mancano indicazioni esplicite nel testo o nel paratesto. I manoscritti datati con certezza (perché il copista ha indicato la data nel colophon, o perché sono databili attraverso riferimenti storici interni) costituiscono solo una piccola frazione del patrimonio manoscritto. Per la maggioranza dei manoscritti, la datazione deve basarsi su criteri indiretti: paleografici (lo stile della scrittura), codicologici (le caratteristiche materiali come il tipo di supporto, la fascicolazione, le dimensioni), testuali (il contenuto e le varianti presenti).

L'approccio bayesiano offre un quadro rigoroso per integrare queste diverse evidenze e produrre stime probabilistiche della data. L'idea fondamentale è di combinare le conoscenze pregresse (prior) con le osservazioni specifiche sul manoscritto (likelihood) per ottenere una distribuzione a posteriori sulla data.

Formalmente, se θ rappresenta la data del manoscritto (una variabile continua, per esempio espressa in anni) e D rappresenta i dati osservati (caratteristiche paleografiche, codicologiche, testuali), il teorema di Bayes ci dice:

$$p(\theta|D) = [p(D|\theta) \times p(\theta)] / p(D)$$

dove $p(\theta|D)$ è la distribuzione a posteriori sulla data, $p(D|\theta)$ è la verosimiglianza dei dati data la data, $p(\theta)$ è la distribuzione a priori sulla data, e $p(D)$ è una costante di normalizzazione.

La distribuzione a priori $p(\theta)$ codifica le nostre conoscenze preliminari prima di esaminare il manoscritto specifico. Se non abbiamo alcuna informazione, possiamo usare un prior uniforme su un intervallo plausibile (per esempio, uniforme tra 800 e 1500 per un manoscritto medievale europeo). Se abbiamo qualche informazione generale (per esempio, sappiamo che il testo fu composto nel XII secolo e che il manoscritto è probabilmente una copia successiva), possiamo usare un prior più informativo, per esempio una distribuzione normale centrata sul XIII secolo con deviazione standard appropriata.

La verosimiglianza $p(D|\theta)$ quantifica quanto sono probabili le osservazioni se il manoscritto risale alla data θ. Questa è la parte più delicata da specificare, perché richiede modelli che mettono in relazione le caratteristiche osservabili con la data. Per le features paleografiche, possiamo utilizzare modelli addestrati su manoscritti datati che predicono la distribuzione delle features in funzione della data. Per esempio, se sappiamo che l'angolo medio della penna nei manoscritti carolina del IX secolo segue una distribuzione $N(35°, 5°)$, mentre nei manoscritti carolina del X secolo segue $N(40°, 5°)$, possiamo calcolare la verosimiglianza di un angolo osservato di 38° sotto le due ipotesi di datazione.

Per caratteristiche codicologiche come il tipo di supporto, possiamo usare informazioni sulla frequenza di utilizzo di diversi materiali nel tempo. Per esempio, l'uso della pergamena è quasi universale fino al XIV secolo, poi viene progressivamente sostituito dalla carta in molte regioni. Se un manoscritto è su carta, la verosimiglianza di una datazione anteriore al 1300 sarà molto bassa in Europa occidentale (mentre potrebbe non esserlo per regioni come l'Italia dove la carta fu introdotta prima).

L'integrazione di evidenze multiple è uno dei punti di forza dell'approccio bayesiano. Se abbiamo m diverse categorie di evidenze (per esempio, paleografia, codicologia, contenuto testuale), e assumiamo che siano condizionatamente indipendenti data la data, la verosimiglianza totale è il prodotto delle verosimiglianze individuali:

$$p(D|\theta) = \prod_{i=1}^{m} p(D_i|\theta)$$

Nella pratica, l'assunzione di indipendenza condizionale è spesso una semplificazione (per esempio, certe caratteristiche paleografiche e codicologiche tendono a co-occorrere), ma fornisce una buona prima approssimazione e rende il calcolo trattabile.

Il risultato dell'analisi bayesiana è una distribuzione a posteriori $p(\theta|D)$, che fornisce non solo una stima puntuale della data (per esempio, la moda o la media della distribuzione) ma anche una misura dell'incertezza (per esempio, un intervallo di credibilità al 95%). Questo è fondamentale perché la datazione paleografica è intrinsecamente incerta, e fornire un singolo anno senza indicazione dell'incertezza è fuorviante.

3.5 Analisi del disegno preparatorio: infrared reflectography

Molti manoscritti miniati presentano disegni preparatori eseguiti prima dell'applicazione dei pigmenti. Questi disegni, realizzati tipicamente con punta metallica, carboncino, o inchiostro diluito, sono spesso invisibili o appena visibili all'occhio nudo una volta che la miniatura è stata completata. La reflectografia infrarossa (infrared reflectography, IRR) è una tecnica di imaging che permette di visualizzare questi disegni preparatori.

Il principio fisico è il seguente: molti pigmenti usati nelle miniature medievali sono relativamente trasparenti alla radiazione infrarossa, mentre i materiali usati per i disegni preparatori (specialmente il carboncino e certi inchiostri) assorbono fortemente l'infrarosso. Illuminando la miniatura con luce infrarossa (tipicamente lunghezze d'onda tra 1000 e 2500 nanometri) e registrando la radiazione riflessa con una camera sensibile all'infrarosso, è possibile "vedere attraverso" gli strati superficiali di pigmento e rivelare il disegno sottostante.

La reflectografia infrarossa ha rivelato informazioni preziose sul processo creativo dei miniatori. Si possono osservare pentimenti (modifiche al disegno originale), correzioni, variazioni nella composizione. In alcuni casi, il disegno preparatorio mostra una complessità e una qualità superiori alla miniatura finale, suggerendo che il disegno era opera di un maestro più abile che poi affidava la colorazione a assistenti meno qualificati.

Dal punto di vista quantitativo, le immagini infrarosse possono essere elaborate per estrarre automaticamente le linee del disegno preparatorio. Algoritmi di edge detection (rilevamento dei bordi) come il filtro di Canny o il filtro di Sobel possono identificare i contorni nel disegno. Tecniche più sofisticate utilizzano reti neurali convoluzionali addestrate specificamente per segmentare i disegni preparatori dalle immagini infrarosse.

Una volta estratte le linee del disegno, è possibile analizzarle quantitativamente. Per esempio, si può studiare lo stile del tratto: linee rapide e sicure suggeriscono un artista esperto, mentre linee esitanti e ripetute indicano minor sicurezza. Si possono confrontare quantitativamente disegni preparatori di diverse miniature per determinare se sono opera della stessa mano, utilizzando metriche di similarità basate su features grafiche.

3.6 Imaging multispettrale e iperspettrale: tecnologie avanzate per l'analisi dei manoscritti

L'imaging multispettrale (Multispectral Imaging, MSI) estende il concetto della reflectografia infrarossa acquisendo immagini del manoscritto a multiple lunghezze d'onda attraverso lo spettro elettromagnetico, dall'ultravioletto (UV, circa 300-400 nm) attraverso il visibile (400-700 nm) fino all'infrarosso vicino e medio (NIR, 700-2500 nm).

Ciascuna banda spettrale fornisce informazioni diverse. L'ultravioletto è particolarmente utile per rivelare correzioni e cancellature, poiché molti pigmenti e inchiostri hanno fluorescenza caratteristica sotto illuminazione UV. Il visibile fornisce l'immagine che corrisponde alla percezione umana. L'infrarosso, come discusso, penetra molti pigmenti rivelando disegni preparatori e testi cancellati.

Un'acquisizione multispettrale tipica potrebbe includere 10-20 bande spettrali. Per ciascuna banda, si acquisisce un'immagine completa del manoscritto utilizzando filtri appropriati o sorgenti di illuminazione monocromatiche. Il risultato è un "data cube" (cubo di dati) tridimensionale, dove le prime due dimensioni rappresentano le coordinate spaziali sulla pagina e la terza dimensione rappresenta la lunghezza d'onda.

L'imaging iperspettrale (Hyperspectral Imaging, HSI) spinge questo concetto all'estremo, acquisendo immagini a decine o centinaia di bande spettrali molto strette (tipicamente 5-10 nm di larghezza). Questo fornisce uno spettro quasi continuo per ciascun pixel dell'immagine, permettendo una caratterizzazione spettroscopica molto dettagliata.

L'analisi dei data cube multispettrali o iperspettrali utilizza tecniche di analisi multivariata. L'obiettivo principale è tipicamente separare e identificare diversi materiali (inchiostri, pigmenti, supporti) sulla base dei loro spettri caratteristici. La tecnica più comune è l'Analisi delle Componenti Principali (Principal Component Analysis, PCA).

La PCA trasforma il data cube in un nuovo sistema di coordinate dove le nuove variabili (le componenti principali) sono combinazioni lineari delle bande spettrali originali, ordinate per varianza decrescente. La prima componente principale cattura la maggior parte della variabilità nei dati (tipicamente legata alle differenze di illuminazione e riflettanza generale), mentre le componenti successive catturano variazioni più sottili, spesso correlate con differenze nei materiali.

Matematicamente, la PCA si basa sulla decomposizione agli autovalori della matrice di covarianza delle bande spettrali. Se X è una matrice n×m dove n è il numero di pixel e m il numero di bande spettrali, la matrice di covarianza è:

$$C = (1/(n-1)) \times X^T \times X$$

Gli autovettori di C (le componenti principali) definiscono le nuove direzioni nello spazio spettrale, e gli autovalori corrispondenti quantificano la varianza lungo ciascuna direzione. Proiettando i dati originali sulle componenti principali:

$$Y = X \times V$$

dove V è la matrice degli autovettori, otteniamo le immagini delle componenti principali. Tipicamente, visualizzando la seconda o terza componente principale, emergono chiaramente distinzioni tra inchiostri diversi o tra testo e supporto che non erano evidenti nelle immagini originali.

Tecniche più sofisticate per l'analisi di dati iperspettrali includono il Minimum Noise Fraction (MNF), simile alla PCA ma che massimizza il rapporto segnale-rumore piuttosto che la varianza, e il Spectral Angle Mapper (SAM), che classifica i pixel basandosi sull'angolo tra il loro spettro e spettri di riferimento noti di materiali puri.

palinsesti , tecnologie di imaging avanzato e recupero di testi cancellati

4.1 palinsesti : definizione e contesto storico

I palinsesti rappresentano una delle categorie più affascinanti e problematiche di manoscritti. Il termine deriva dal greco παλίμψηστον (palímpsēston), da πάλιν (pálin, "di nuovo") e ψάω (psáō, "raschio"), e designa un supporto scrittorio, tipicamente pergamena, da cui il testo originale è stato cancellato per permettere il riutilizzo del supporto per un nuovo testo.

La pratica del riutilizzo della pergamena mediante raschiatura o lavaggio del testo preesistente è documentata fin dall'antichità, ma divenne particolarmente comune in certi periodi storici caratterizzati da scarsità di materiali scrittori o da cambiamenti culturali che rendevano certi testi "obsoleti" o meno preziosi. Il periodo più significativo per la produzione di palinsesti in Europa fu l'alto Medioevo (VI-IX secolo), quando la transizione dall'antichità classica al Medioevo cristiano portò alla cancellazione di numerosi testi classici pagani per far posto a testi liturgici o patristici.

La pergamena, prodotta dalla pelle di animali (pecore, capre, vitelli) mediante un processo laborioso di pulitura, raschiatura, stretching e sbiancamento, era un materiale costoso. Un singolo codice di dimensioni medie richiedeva le pelli di decine di animali. In contesti di ristrettezze economiche o di urgente necessità di supporti scrittori (per esempio, per la produzione di libri liturgici necessari al culto), il riutilizzo di pergamena da manoscritti considerati meno essenziali era una soluzione pragmatica.

Il processo di cancellazione poteva avvenire in diversi modi. Il metodo più radicale era la raschiatura meccanica: la superficie della pergamena veniva abrasa con pomice o altri materiali abrasivi per rimuovere fisicamente lo strato superficiale contenente l'inchiostro. Un metodo alternativo era il lavaggio chimico: la pergamena veniva trattata con sostanze (per esempio, latte, acidi deboli, soluzioni alcaline) che dissolvevano o scolorivano l'inchiostro. Spesso si utilizzava una combinazione di entrambi i metodi.

Nonostante questi processi di cancellazione, in molti casi il testo originale non veniva completamente obliterato. Tracce di inchiostro rimanevano intrappolate nelle fibre della pergamena o negli strati più profondi della pelle. Queste tracce, invisibili o appena percettibili all'occhio nudo, possono essere rivelate mediante tecnologie di imaging avanzate, permettendo il recupero di testi altrimenti perduti.

4.2 Il Palinsesto di Archimede: un caso esemplare

Il Palinsesto di Archimede costituisce uno dei casi più celebri e scientificamente significativi di recupero di un testo da un palimpsesto. Il manoscritto, noto anche come Codex C, è un euchologion (libro di preghiere) bizantino copiato nel XII-XIII secolo. La pergamena utilizzata per questo libro liturgico proveniva però da almeno sette manoscritti più antichi che erano stati cancellati per permetterne il riutilizzo. Uno di questi manoscritti più antichi conteneva opere di Archimede di Siracusa, il grande matematico e fisico greco del III secolo a.C., copiate in greco antico nel X secolo.

La storia della riscoperta è avventurosa quanto il contenuto del manoscritto. Il palimpsesto fu identificato per la prima volta dal filologo danese Johan Ludvig Heiberg nel 1906 nella biblioteca del monastero di San Saba a Costantinopoli (l'odierna Istanbul). Heiberg riuscì a

decifrare parti del testo sottostante mediante esame diretto e fotografie, riconoscendo la presenza di opere di Archimede, alcune note e altre fino ad allora sconosciute. Dopo complesse vicende che includono la dispersione durante le guerre balcaniche, il passaggio in collezioni private, e controversie legali, il manoscritto fu acquisito da un anonimo collezionista nel 1998 e affidato al Walters Art Museum di Baltimora per lo studio e la conservazione.

Le tecnologie di imaging applicate al Palinsesto di Archimede tra il 1999 e il 2008 rappresentano un milestone nella codicologia digitale. Fu utilizzata una combinazione di tecniche: fotografia multispettrale (con illuminazione UV, visibile in varie bande, e infrarosso), reflectografia infrarossa, fluorescenza a raggi X, e imaging iperspettrale. Queste tecnologie hanno permesso di leggere circa il 90% del testo archimedeo sottostante, nonostante le difficoltà create dalla cancellazione originale e dal deterioramento successivo.

Il contenuto matematico rivelato è di straordinaria importanza. Il palimpsesto contiene l'unica copia conosciuta del trattato di Archimede "Metodo dei teoremi meccanici" (Ephodos), in cui Archimede spiega come utilizzò ragionamenti di meccanica e considerazioni infinitesimali per scoprire molti dei suoi risultati geometrici prima di dimostrarli rigorosamente. Questo testo illumina il processo euristico di uno dei più grandi matematici della storia.

Ancora più notevole è la presenza nel palimpsesto dell'unica copia conosciuta dello "Stomachion", un trattato che tratta di un antico puzzle combinatorio. La decifrazione di questo testo ha rivelato che Archimede aveva sviluppato approcci alla combinatoria con quasi due millenni di anticipo rispetto allo sviluppo moderno di questo campo matematico. Il testo contiene enumerazioni e considerazioni che anticipano concetti della moderna teoria combinatoria.

4.3 Tecnologie di imaging per palinsesti : principi fisici

Il successo nel recupero di testi da palinsesti dipende crucialmente dalla comprensione dei principi fisici dell'interazione tra radiazione elettromagnetica e materiali. I diversi tipi di imaging sfruttano differenti fenomeni fisici per massimizzare il contrasto tra il testo cancellato e il supporto o il testo sovrascritto.

L'imaging ultravioletto sfrutta la fluorescenza. Quando materiali organici (inclusi molti inchiostri e la pergamena stessa) sono illuminati con luce UV, assorbono fotoni ad alta energia e riemettono fotoni a energie più basse nel visibile. Diversi materiali hanno spettri di fluorescenza caratteristici. Gli inchiostri ferro-gallici, molto comuni nei manoscritti medievali, mostrano tipicamente una fluorescenza debole o assente, apparendo scuri contro la fluorescenza più intensa della pergamena. Questo contrasto può rivelare tracce di inchiostro cancellato che non sono visibili in luce normale.

La quantificazione della fluorescenza UV può essere modellata mediante la legge di Beer-Lambert modificata per la fluorescenza. Se I_0 è l'intensità della radiazione UV incidente, la frazione assorbita dal materiale a profondità x è:

$dI/dx = -\mu(\lambda_{ex}) \times I(x)$

dove μ(λ_ex) è il coefficiente di assorbimento alla lunghezza d'onda di eccitazione λ_ex. L'intensità della fluorescenza emessa a lunghezza d'onda λ_em è proporzionale alla radiazione assorbita e al rendimento quantico di fluorescenza Φ:

$I_{fl}(\lambda_{em}) \propto \Phi(\lambda_{em}) \times \int \mu(\lambda_{ex}) \times I(x)\, dx$

Materiali con alto coefficiente di assorbimento UV e alto rendimento quantico appariranno brillanti in imaging di fluorescenza UV, mentre materiali con basso rendimento quantico appariranno scuri.

L'imaging infrarosso, come discusso precedentemente, sfrutta la trasparenza differenziale dei materiali alle diverse lunghezze d'onda. Nel caso dei palinsesti , l'obiettivo è penetrare il testo sovrascritto per rivelare quello sottostante. Questo è possibile quando l'inchiostro del testo sovrascritto è relativamente trasparente all'infrarosso mentre quello del testo cancellato assorbe fortemente. Gli inchiostri al carbonio (nerofumo, lampblack), per esempio, assorbono fortemente attraverso tutto lo spettro dal visibile all'infrarosso, mentre molti inchiostri ferro-gallici diventano più trasparenti a lunghezze d'onda maggiori di 1000 nm.

La profondità di penetrazione della radiazione nel materiale è governata dal coefficiente di estinzione:

$I(z) = I_0 \times \exp(-\alpha \times z)$

dove α è il coefficiente di estinzione (che include sia assorbimento sia scattering) e z è la profondità. Per massimizzare la penetrazione e quindi la visibilità del testo sottostante, si scelgono lunghezze d'onda per cui α del testo sovrascritto è minimo.

4.4 Fluorescenza a raggi X: analisi elementale non distruttiva

Una tecnica particolarmente potente per l'analisi dei palinsesti è la fluorescenza a raggi X (X-Ray Fluorescence, XRF). Questa tecnica permette di identificare gli elementi chimici presenti negli inchiostri mediante eccitazione con raggi X.

Il principio fisico è il seguente. Quando un materiale è bombardato con raggi X sufficientemente energetici, gli elettroni degli orbitali più interni degli atomi vengono espulsi (effetto fotoelettrico). Gli orbitali più esterni hanno energia maggiore, e quando un elettrone da un orbitale esterno "decade" per riempire la vacanza creata nell'orbitale interno, l'energia in eccesso viene emessa sotto forma di un fotone X caratteristico. L'energia di questi fotoni è specifica per ciascun elemento, costituendo una "firma spettroscopica" che permette l'identificazione.

Per esempio, il ferro (Fe), componente principale degli inchiostri ferro-gallici, ha linee di fluorescenza caratteristiche K_α a 6,40 keV e K_β a 7,06 keV. Il rame (Cu), talvolta presente negli inchiostri verdi o blu, ha linee K_α a 8,05 keV e K_β a 8,91 keV. Il calcio (Ca), presente nella pergamena per trattamenti di preparazione, ha linee K_α a 3,69 keV.

L'imaging XRF (X-Ray Fluorescence Imaging) estende questa analisi puntuale a un'immagine completa. Scansionando il fascio di raggi X attraverso il manoscritto e registrando lo spettro di fluorescenza in ciascun punto, si ottiene un data cube

tridimensionale: due dimensioni spaziali e una dimensione spettrale (energia dei fotoni X). Selezionando l'energia corrispondente a un elemento specifico, si può produrre una mappa della distribuzione spaziale di quell'elemento.

Nel contesto dei palinsesti , l'imaging XRF è particolarmente utile quando il testo cancellato e quello sovrascritto utilizzano inchiostri con composizioni chimiche diverse. Per esempio, se il testo sottostante era scritto con inchiostro ferro-gallico (ricco di Fe) e quello sovrascritto con inchiostro al carbonio (contenente principalmente C, non rilevabile con XRF), si può produrre una mappa della distribuzione del ferro che rivelerà selettivamente il testo cancellato.

Il Palinsesto di Archimede fu uno dei primi manoscritti ad essere analizzato sistematicamente con imaging XRF. Le mappe di distribuzione del ferro hanno rivelato dettagli del testo archimedeo che erano invisibili con altre tecniche, inclusi diagrammi geometrici che erano stati quasi completamente cancellati. La risoluzione spaziale dell'imaging XRF (tipicamente 50-100 micrometri con i sistemi moderni) è sufficiente per distinguere singole lettere.

4.5 Imaging iperspettrale: separazione di inchiostri sovrapposti

Quando un palimpsesto presenta testo cancellato e testo sovrascritto scritti con inchiostri spettralmente simili, le tecniche di imaging a singola banda possono non essere sufficienti per separare i due strati. L'imaging iperspettrale, acquisendo uno spettro completo per ciascun pixel, fornisce informazioni molto più ricche che possono essere sfruttate per la separazione.

Il problema può essere formulato come unmixing spettrale (separazione di miscele spettrali). In ciascun pixel, lo spettro osservato è una combinazione degli spettri dei materiali puri presenti:

$S(\lambda) = \Sigma_i a_i \times E_i(\lambda) + n(\lambda)$

dove $S(\lambda)$ è lo spettro osservato alla lunghezza d'onda λ, $E_i(\lambda)$ sono gli spettri degli "endmembers" (materiali puri, come inchiostro del testo sottostante, inchiostro del testo sovrascritto, pergamena), a_i sono i coefficienti di abbondanza (quanto di ciascun materiale è presente nel pixel), e $n(\lambda)$ rappresenta il rumore.

L'obiettivo dell'unmixing spettrale è stimare gli endmember E_i e le abbondanze a_i a partire dagli spettri osservati. Esistono due approcci principali: linear unmixing (basato sul modello lineare sopra) e nonlinear unmixing (che tiene conto di fenomeni di scattering multiplo che rendono la miscela non lineare).

Per il linear unmixing, il metodo più comune è la Nonnegative Matrix Factorization (NMF), che decompone la matrice dei dati spettrali X (ciascuna riga è lo spettro di un pixel) come prodotto di due matrici non negative:

$X \approx W \times H$

dove W contiene gli spettri degli endmember (una colonna per ciascun endmember) e H contiene le abbondanze (una riga per ciascun endmember, una colonna per ciascun pixel). Il vincolo di non negatività è fisicamente motivato: né gli spettri né le abbondanze possono essere negativi.

La fattorizzazione si ottiene minimizzando una funzione obiettivo che misura la discrepanza tra X e W×H, tipicamente la norma di Frobenius al quadrato:

min_(W,H≥0) ||X - W×H||_F^2

Questo problema di ottimizzazione non convesso si risolve iterativamente mediante algoritmi come Multiplicative Update Rules o Alternating Least Squares.

Una volta ottenuti gli endmember e le abbondanze, si possono produrre immagini che mostrano selettivamente ciascun componente. Per esempio, visualizzando l'abbondanza dell'endmember corrispondente all'inchiostro del testo cancellato, si ottiene un'immagine che evidenzia quel testo sopprimendo il testo sovrascritto.

Nel caso del Palinsesto di Archimede, l'imaging iperspettrale con 50 bande spettrali tra 365 e 950 nm, seguito da analisi PCA e unmixing spettrale, ha permesso di separare il testo archimedeo dal testo liturgico sovrascritto e da falsi del XX secolo (il manoscritto era stato oggetto di manipolazioni e aggiunte fraudolente). La separazione ha raggiunto una qualità tale che il testo archimedeo poteva essere letto quasi come se il testo sovrascritto non esistesse.

4.6 Tecniche di elaborazione delle immagini per il miglioramento della leggibilità

Una volta acquisite le immagini mediante le varie modalità di imaging, il passo successivo è l'elaborazione digitale per massimizzare la leggibilità del testo. Questo coinvolge una serie di operazioni di image processing.

Il contrast enhancement (miglioramento del contrasto) è l'operazione più fondamentale. Il contrasto tra testo e sfondo è spesso basso nelle immagini di palinsesti , rendendo difficile la lettura. Tecniche semplici come lo stretching dell'istogramma (riscalare linearmente l'intervallo di intensità per utilizzare l'intera gamma dinamica disponibile) o l'equalizzazione dell'istogramma (ridistribuire le intensità per ottenere un istogramma uniforme) possono migliorare significativamente la visibilità.

Per immagini con illuminazione non uniforme (shading), tecniche più sofisticate sono necessarie. Un approccio comune è la correzione mediante modellazione dello sfondo: si stima la componente di illuminazione I_bg(x,y) (tipicamente mediante filtri passa-basso o fitting di superfici), poi si normalizza l'immagine:

I_corr(x,y) = I_orig(x,y) / I_bg(x,y)

Questo rimuove le variazioni lente di intensità dovute all'illuminazione mantenendo le variazioni rapide (i bordi delle lettere).

Il denoising (riduzione del rumore) è critico quando le immagini sono acquisite in condizioni di basso segnale (per esempio, imaging di fluorescenza debole). Filtri adattivi come il filtro di Wiener, che stima il segnale originale minimizzando l'errore quadratico medio considerando le statistiche locali del rumore, sono molto efficaci. Metodi più recenti basati su non-local means (sfruttando la similarità tra patch dell'immagine) o su sparse representations (rappresentazioni sparse) offrono prestazioni superiori.

Per palinsesti particolarmente problematici, dove testo cancellato e testo sovrascritto si sovrappongono, tecniche di separazione di fonti possono essere applicate. Independent Component Analysis (ICA), che cerca proiezioni dei dati che siano statisticamente indipendenti, può separare testo e sfondo assumendo che siano sorgenti indipendenti. La formulazione matematica è:

$X = A \times S$

dove X sono i dati osservati (immagini multispettrali), S sono le sorgenti indipendenti (immagini dei componenti separati), e A è la matrice di miscelazione. L'ICA stima A e S imponendo l'indipendenza statistica delle righe di S, misurata per esempio mediante la minimizzazione della mutual information.

4.7 Ricostruzione 3D e analisi della superficie della pergamena

Tecniche emergenti di imaging tridimensionale permettono di analizzare la microtopografia della superficie della pergamena, rivelando informazioni sul processo di raschiatura e sulle tracce meccaniche lasciate dal testo cancellato.

Il Polynomial Texture Mapping (PTM) è una tecnica che acquisisce immagini dell'oggetto sotto illuminazioni da diverse direzioni, poi modella la riflettanza di ciascun pixel come funzione della direzione di illuminazione. Questo permette di riilluminare virtualmente l'oggetto da qualsiasi direzione, evidenziando dettagli superficiali che dipendono dalle variazioni di rilievo microscopico.

Matematicamente, per ciascun pixel, la luminanza L è modellata come funzione polinomiale delle componenti (u,v) della direzione di illuminazione:

$L(u,v) = a_0 \times u^2 + a_1 \times v^2 + a_2 \times u \times v + a_3 \times u + a_4 \times v + a_5$

I sei coefficienti a_i si stimano mediante least squares fitting dalle immagini acquisite sotto almeno sei diverse direzioni di illuminazione. Una volta ottenuto il PTM, si può sintetizzare l'immagine per qualsiasi direzione di illuminazione, includendo direzioni radenti che massimizzano le ombre e quindi il contrasto delle irregolarità superficiali.

Nel caso di palinsesti , anche dopo la raschiatura, il processo di scrittura lascia spesso impronte microscopiche nella pergamena (compressione delle fibre sotto la pressione dello stilo o della penna). Il PTM può rivelare queste tracce che non sono visibili con illuminazione normale, permettendo di leggere testo che è stato completamente privato dell'inchiostro.

Survival analysis, degradazione dei materiali e modelli cinetici del deterioramento

5.1 Survival analysis applicata ai manoscritti: quanto durano i codici?

Una domanda fondamentale per la codicologia e per la comprensione della trasmissione culturale è: quanto a lungo sopravvivono i manoscritti? E quali fattori influenzano la probabilità di sopravvivenza? L'analisi di sopravvivenza (survival analysis), una branca della statistica sviluppata originariamente per l'epidemiologia e la biostatistica, fornisce gli strumenti concettuali e metodologici per affrontare queste domande.

In termini formali, definiamo la variabile T come il tempo di sopravvivenza di un manoscritto, ossia il tempo che intercorre tra la produzione del manoscritto e la sua distruzione o perdita definitiva. La funzione di sopravvivenza S(t) rappresenta la probabilità che un manoscritto sopravviva oltre il tempo t:

$S(t) = P(T > t)$

La funzione di sopravvivenza è monotona decrescente ($S(0) = 1$, $S(\infty) = 0$) e caratterizza completamente la distribuzione dei tempi di sopravvivenza.

Una funzione strettamente correlata è la funzione di hazard h(t), che rappresenta il tasso istantaneo di fallimento (nel nostro caso, distruzione o perdita del manoscritto) al tempo t, condizionato alla sopravvivenza fino a quel momento:

$h(t) = \lim_{\Delta t \to 0} P(t < T \leq t+\Delta t \mid T > t) / \Delta t$

La funzione di hazard può essere interpretata come il "rischio" di perdita in funzione dell'età del manoscritto. La relazione tra funzione di sopravvivenza e funzione di hazard è:

$S(t) = \exp[-\int_0^t h(s)\, ds]$

Questa relazione permette di calcolare la funzione di sopravvivenza se conosciamo la funzione di hazard, e viceversa.

5.2 Distribuzione di Weibull: modellazione parametrica della sopravvivenza

Una delle distribuzioni più utilizzate per modellare tempi di sopravvivenza in contesti tecnici e biologici è la distribuzione di Weibull. Questa distribuzione è particolarmente flessibile e può rappresentare diversi pattern di hazard (costante, crescente, decrescente) a seconda dei valori dei parametri.

La distribuzione di Weibull ha due parametri: il parametro di scala λ (che determina la scala temporale caratteristica) e il parametro di forma k (che determina come il rischio varia nel tempo). La funzione di densità di probabilità è:

$f(t) = (k/\lambda) \times (t/\lambda)^{(k-1)} \times \exp[-(t/\lambda)^k]$

per t ≥ 0. La funzione di sopravvivenza corrispondente è:

$S(t) = \exp[-(t/\lambda)^k]$

e la funzione di hazard è:

$h(t) = (k/\lambda) \times (t/\lambda)^{(k-1)}$

Il parametro di forma k determina il comportamento qualitativo della funzione di hazard:

- Se $k < 1$, la funzione di hazard è decrescente nel tempo: il rischio di perdita diminuisce con l'età del manoscritto. Questo potrebbe riflettere una selezione iniziale dove i manoscritti più fragili o meno apprezzati vengono perduti rapidamente, mentre quelli che sopravvivono oltre una certa età hanno buone probabilità di sopravvivere ancora a lungo.
- Se $k = 1$, la funzione di hazard è costante (distribuzione esponenziale): il rischio di perdita non dipende dall'età. Questo corrisponderebbe a perdite casuali, indipendenti dalla qualità o dall'età del manoscritto.
- Se $k > 1$, la funzione di hazard è crescente: il rischio aumenta con l'età, riflettendo processi di invecchiamento e degradazione progressiva dei materiali.

Gli studi empirici sulla sopravvivenza dei manoscritti medievali suggeriscono valori di k tipicamente maggiori di 1 ma non molto elevati ($k \approx 1{,}2\text{-}1{,}8$), indicando un moderato aumento del rischio con l'età, compatibile con processi lenti di degradazione fisica e chimica.

Il parametro di scala λ fornisce una misura caratteristica del tempo di sopravvivenza. Per la distribuzione di Weibull, la vita media (tempo medio di sopravvivenza) è:

$E[T] = \lambda \times \Gamma(1 + 1/k)$

dove Γ è la funzione gamma. La mediana del tempo di sopravvivenza (il tempo entro cui il 50% dei manoscritti è perduto) è:

$t_{mediana} = \lambda \times [\ln(2)]^{(1/k)}$

Per fare un esempio concreto, supponiamo che l'analisi statistica di un corpus di manoscritti medievali produca stime $\lambda = 800$ anni e $k = 1{,}5$. Questo implica una vita media di circa $800 \times \Gamma(1{,}667) \approx 800 \times 0{,}903 \approx 722$ anni, e una mediana di circa $800 \times [0{,}693]^{(1/1{,}5)} \approx 800 \times 0{,}785 \approx 628$ anni. La funzione di sopravvivenza sarebbe $S(1000) = \exp[-(1000/800)^{1{,}5}] \approx 0{,}28$, indicando che circa il 28% dei manoscritti sopravvive oltre 1000 anni.

5.3 Funzioni di hazard e fattori di rischio: analisi di Cox

Nella realtà, il rischio di perdita di un manoscritto non dipende solo dall'età ma anche da numerosi altri fattori: il materiale di supporto (pergamena vs carta), la qualità della manifattura, il contenuto (testi liturgici vs opere profane), le condizioni di conservazione (biblioteche monastiche vs proprietà privata), eventi storici (guerre, incendi, dispersioni).

Il modello di regressione di Cox (Cox proportional hazards model) fornisce un quadro flessibile per analizzare l'effetto di questi fattori (chiamati covariates o predittori) sulla

funzione di hazard, senza richiedere di specificare la forma funzionale della dipendenza temporale di base.

Il modello di Cox assume che la funzione di hazard per un individuo (manoscritto) con covariates x = (x_1, x_2, ..., x_p) sia:

$h(t|x) = h_0(t) \times \exp(\beta_1 \times x_1 + \beta_2 \times x_2 + ... + \beta_p \times x_p)$

dove $h_0(t)$ è la baseline hazard function (la funzione di hazard per un individuo con tutte le covariates uguali a zero), e i coefficienti β_i quantificano l'effetto delle covariates. La forma esponenziale assicura che la funzione di hazard sia sempre positiva.

Il termine "proportional hazards" deriva dal fatto che il rapporto tra le funzioni di hazard di due individui con covariates diverse è costante nel tempo:

$h(t|x')/h(t|x) = \exp[\beta \times (x' - x)]$

non dipende da t. Questo significa che una covariate che aumenta il rischio lo fa proporzionalmente, indipendentemente dall'età del manoscritto.

I coefficienti β si stimano mediante massimizzazione della partial likelihood, una quantità che dipende solo dall'ordine in cui gli eventi (perdite di manoscritti) si verificano, non dai tempi assoluti. Questo è un vantaggio notevole perché permette l'inferenza anche quando i tempi esatti di evento non sono conosciuti con precisione, come spesso accade per i manoscritti storici.

L'interpretazione dei coefficienti è in termini di hazard ratio. Per una covariate binaria (per esempio, $x_1 = 1$ se il manoscritto è su carta, $x_1 = 0$ se su pergamena), il coefficiente β_1 corrisponde al logaritmo del hazard ratio:

$HR = \exp(\beta_1) = h(t|x_1=1) / h(t|x_1=0)$

Se HR = 2, significa che i manoscritti su carta hanno un rischio di perdita doppio rispetto a quelli su pergamena in ogni momento, a parità di altre condizioni. Se $\beta_1 = 0{,}693$ ($\ln(2)$), allora HR = 2.

Applicazioni del modello di Cox ai manoscritti medievali hanno rivelato diversi fattori significativi per la sopravvivenza. Il materiale di supporto è cruciale: i manoscritti su pergamena hanno tassi di sopravvivenza significativamente superiori a quelli su carta, con hazard ratio tipicamente HR ≈ 0,3-0,5 (ossia, il rischio di perdita è ridotto del 50-70%). Il contenuto è anch'esso importante: manoscritti liturgici e testi biblici hanno tassi di sopravvivenza più elevati rispetto a opere profane di minor valore cultuale. La provenienza istituzionale conta: manoscritti conservati continuativamente in biblioteche monastiche o universitarie hanno mostrato tassi di sopravvivenza molto superiori rispetto a quelli passati attraverso collezioni private o mercato antiquario.

5.4 Degradazione dei materiali: processi chimici e cinetici

La sopravvivenza fisica dei manoscritti è limitata da processi di degradazione chimica e fisica dei materiali costituenti: pergamena, carta, inchiostri, pigmenti, legature. La

comprensione di questi processi è fondamentale sia per spiegare i pattern di sopravvivenza osservati sia per sviluppare strategie di conservazione efficaci.

La pergamena, costituita principalmente da collagene (una proteina fibrosa), degrada attraverso due processi principali: idrolisi e ossidazione. L'idrolisi acida o basica scinde i legami peptidici nella catena proteica, riducendo il peso molecolare medio e compromettendo le proprietà meccaniche. L'ossidazione, catalizzata da metalli di transizione (ferro, rame) e accelerata dalla luce, danneggia specifici amminoacidi (specialmente quelli contenenti zolfo come cisteina e metionina) e può causare crosslinking intermolecolare che irrigidisce e fragilizza la pergamena.

La cinetica dell'idrolisi del collagene segue approssimativamente una legge di primo ordine:

$$d[P]/dt = -k \times [P]$$

dove [P] è la concentrazione di legami peptidici intatti e k è la costante di velocità. La soluzione è:

$$P = [P]_0 \times \exp(-k \times t)$$

La costante di velocità k dipende fortemente dalla temperatura secondo l'equazione di Arrhenius:

$$k(T) = A \times \exp(-E_a / (R \times T))$$

dove A è il fattore pre-esponenziale, E_a è l'energia di attivazione (tipicamente 80-120 kJ/mol per l'idrolisi del collagene), R è la costante dei gas (8,314 J/(mol×K)), e T è la temperatura assoluta in Kelvin.

Questa dipendenza dalla temperatura implica che la conservazione a temperatura ridotta rallenta drammaticamente la degradazione. Una riduzione di temperatura da 25°C (298 K) a 15°C (288 K) comporta una riduzione del tasso di degradazione di un fattore:

$$k(288)/k(298) = \exp[-E_a/R \times (1/288 - 1/298)] \approx \exp[-100000/8{,}314 \times (-0{,}000116)] \approx \exp(1{,}40) \approx 4{,}05$$

ossia la degradazione procede circa 4 volte più lentamente a 15°C rispetto a 25°C. Questo spiega perché le biblioteche mantengono temperature controllate relativamente basse.

L'umidità relativa (RH) gioca anch'essa un ruolo critico. Valori di RH troppo bassi (< 30%) causano essiccazione e fragilizzazione della pergamena, mentre valori troppo alti (> 65%) favoriscono la crescita fungina e accelerano l'idrolisi. Il range ottimale è tipicamente 45-55% RH.

La carta, costituita da fibre di cellulosa, degrada principalmente attraverso idrolisi acido-catalizzata. Molte carte storiche prodotte dal XVIII secolo in poi sono acide (pH < 7) a causa dell'uso di allume nel processo di collatura, e subiscono auto-catalisi: i prodotti di degradazione (acidi organici) accelerano ulteriormente la degradazione. Questo fenomeno è particolarmente grave per le carte prodotte con pasta di legno contenente lignina, che si ossida producendo ulteriori specie acide.

La cinetica di degradazione della carta acida è stata oggetto di numerosi studi. Il grado di polimerizzazione (DP, numero medio di unità di glucosio nelle catene di cellulosa) diminuisce secondo una legge empirica simile a quella del collagene:

$$DP(t) = DP_0 \times \exp(-k \times t)$$

con costanti di velocità che dipendono dal pH e dalla temperatura. Per carta fortemente acida (pH ≈ 4-5), i tempi di dimezzamento del DP a temperatura ambiente sono dell'ordine di decenni, spiegando il deterioramento rapido di molti libri del XIX e XX secolo rispetto ai manoscritti medievali su pergamena.

5.5 Attacco fungino e biologico: modelli di crescita e danni

Una minaccia particolarmente insidiosa per i manoscritti è l'attacco da parte di microrganismi, specialmente funghi (muffe). I funghi possono crescere sulla superficie o all'interno dei materiali quando le condizioni di umidità e temperatura sono favorevoli, producendo enzimi (cellulasi, proteasi) che degradano cellulosa e proteine, causando macchie, scolorimenti, perdita di resistenza meccanica, e nei casi gravi disgregazione completa del supporto.

La crescita fungina può essere modellata mediante equazioni logistiche. Se N rappresenta la biomassa fungina (per esempio, in unità arbitrarie di densità di micelio), la crescita segue:

$$dN/dt = r \times N \times (1 - N/K)$$

dove r è il tasso di crescita intrinseco e K è la capacità portante (carrying capacity, la massima biomassa sostenibile). La soluzione è la curva logistica:

$$N(t) = K / [1 + (K/N_0 - 1) \times \exp(-r \times t)]$$

che mostra crescita inizialmente esponenziale (quando $N << K$) seguita da rallentamento e saturazione al valore K.

Il tasso di crescita r dipende criticamente dalla temperatura e dall'umidità relativa. La maggior parte dei funghi che attaccano i materiali librari (generi come Aspergillus, Penicillium, Cladosporium) hanno tassi di crescita massimi a temperature di 25-30°C e umidità relativa di 75-95%. Al di sotto di circa 65% RH, la crescita è fortemente inibita per la maggior parte delle specie, da cui la raccomandazione di mantenere l'umidità sotto questa soglia critica.

Il danno causato dai funghi può essere quantificato in termini di perdita di resistenza meccanica del materiale. Esperimenti hanno mostrato che la resistenza alla trazione della carta o pergamena colonizzata da funghi diminuisce approssimativamente in proporzione al logaritmo della biomassa fungina:

$$\sigma(N) = \sigma_0 - \beta \times \ln(N/N_0)$$

dove σ è la resistenza alla trazione, σ_0 è la resistenza del materiale non colonizzato, e β è un coefficiente che dipende dal tipo di fungo e materiale. Questo implica che anche colonizzazioni relativamente limitate possono causare perdite significative di resistenza.

Oltre ai funghi, altri agenti biologici di deterioramento includono insetti (specialmente l'Anobium punctatum, il tarlo dei libri, e varie specie di tarme e pesciolini d'argento) e batteri. Gli insetti causano danni meccanici diretti scavando gallerie attraverso carta e pergamena per nutrirsi della cellulosa o delle colle proteiche, e producendo escrementi acidi che accelerano la degradazione chimica. I batteri, in condizioni di alta umidità, possono decomporre materiali organici producendo acidi e pigmenti colorati.

5.6 Photooxidation e fotodegradazione: effetti della luce

L'esposizione alla luce, specialmente alla componente ultravioletta e violetta dello spettro ($\lambda < 450$ nm), causa degradazione fotochimica dei materiali costituenti i manoscritti. I fotoni sufficientemente energetici possono scindere legami chimici, generare radicali liberi reattivi, e innescare reazioni di ossidazione a catena.

La cellulosa, il collagene, e molti inchiostri e pigmenti sono suscettibili alla fotodegradazione. Per la cellulosa, l'assorbimento di fotoni UV porta alla formazione di radicali che attaccano i legami glicosidici, riducendo il grado di polimerizzazione. Il processo è autocatalitico poiché i prodotti di degradazione (gruppi carbonilici) assorbono ancora più fortemente nella regione UV-visibile, accelerando ulteriormente la degradazione.

La cinetica della fotodegradazione è tipicamente descritta mediante una legge di ordine zero o pseudo-primo ordine in funzione della dose di radiazione (intensità × tempo):

$$d(DP)/d(dose) = -k_{photo}$$

dove k_{photo} è un coefficiente di fotodegradazione che dipende dalla lunghezza d'onda della radiazione e dalla presenza di sensibilizzatori (composti che assorbono la luce e trasferiscono l'energia a substrati vicini).

L'integrazione fornisce:

$$DP(dose) = DP_0 - k_{photo} \times dose$$

La dose di radiazione può essere espressa in lux-ore (lux × ore) o meglio in unità radiometriche come joule per metro quadro (J/m^2). Gli studi hanno determinato che per la carta, dosi dell'ordine di 10^7 - 10^8 lux-ora causano dimezzamento delle proprietà meccaniche. Se un manoscritto è esposto a 50 lux (un livello di illuminazione basso, raccomandato per materiali sensibili) per 8 ore al giorno, accumula circa 150.000 lux-ora all'anno. A questo tasso, ci vorrebbero circa 67-670 anni per raggiungere dosi critiche, ma con illuminazioni più intense (300-500 lux, tipiche di esposizioni museali tradizionali) i tempi si riducono a 10-100 anni.

Per minimizzare la fotodegradazione, le raccomandazioni internazionali per la conservazione prescrivono:

- Illuminazione a livelli bassi (50 lux o meno per materiali molto sensibili, 150-200 lux per materiali meno sensibili)
- Filtrazione della componente UV (< 400 nm) mediante filtri o vetri speciali
- Limitazione del tempo di esposizione, specialmente per esposizioni temporanee

- Uso di illuminazione LED che emette poco UV rispetto alle lampade incandescenti o fluorescenti tradizionali

5.7 Modelli integrati di deterioramento: approccio multifisico

I processi di degradazione discussi sopra non operano isolatamente ma interagiscono in modi complessi. Per esempio, la fotodegradazione produce specie chimiche reattive che accelerano l'idrolisi; l'attacco fungino crea microfratture che facilitano la penetrazione di inquinanti atmosferici; l'ossidazione dei componenti della pergamena può catalizzare ulteriore degradazione.

Modelli integrati di deterioramento cercano di catturare queste interazioni. Un approccio è quello dei modelli compartimentali, dove il materiale è suddiviso in diversi "compartimenti" che rappresentano stati di degradazione progressivi (per esempio: intatto → parzialmente degradato → severamente degradato → completamente degradato), e si modellano le transizioni tra compartimenti come processi cinetici con tassi dipendenti da temperatura, umidità, esposizione alla luce, e altri fattori ambientali.

Formalmente, se N_i(t) rappresenta la frazione del materiale nel compartimento i al tempo t, il sistema è descritto da equazioni differenziali accoppiate:

dN_i/dt = Σ_j k_ji × N_j - Σ_j k_ij × N_i

dove k_ij è il tasso di transizione dal compartimento i al compartimento j. Le costanti cinetiche dipendono dalle condizioni ambientali secondo relazioni tipo Arrhenius per la temperatura e relazioni empiriche per l'umidità e altri fattori.

Questi modelli possono essere calibrati su dati di invecchiamento accelerato (esperimenti in cui campioni sono esposti a condizioni estreme di temperatura, umidità, luce per accelerare la degradazione e permettere osservazioni in tempi ragionevoli) ed estrapolati a condizioni di conservazione normali per predire la vita residua dei materiali.

Una sfida importante è la validazione di queste estrapolazioni, poiché i processi accelerati potrebbero non replicare fedelmente i meccanismi che operano a lungo termine in condizioni normali. Confronti con manoscritti storici di età nota forniscono una validazione parziale, sebbene le condizioni di conservazione passate siano spesso incerte.

Digital humanities, collazione automatica e attribuzione d'autore mediante stilometria

6.1 Digital humanities: la rivoluzione digitale nella ricerca umanistica

Le digital humanities (umanistica digitale) rappresentano l'applicazione sistematica di metodi computazionali, tecnologie digitali e approcci quantitativi alla ricerca nelle discipline umanistiche. Nel contesto della critica testuale e della codicologia, le digital humanities hanno rivoluzionato pratiche consolidate da secoli, rendendo possibile l'analisi di corpora testuali di dimensioni senza precedenti, la collaborazione internazionale su edizioni critiche

digitali, e l'applicazione di metodi statistici sofisticati a problemi tradizionalmente affrontati con approcci qualitativi.

Un elemento fondamentale dell'umanistica digitale è la codifica testuale standardizzata. Il consorzio Text Encoding Initiative (TEI) ha sviluppato uno standard basato su XML (eXtensible Markup Language) per la rappresentazione digitale di testi di interesse umanistico. Il TEI-XML permette di codificare non solo il contenuto testuale in sé, ma anche metadati ricchissimi che catturano aspetti filologici, strutturali, semantici e interpretativi del testo.

Un documento TEI-XML è strutturato gerarchicamente in elementi, ciascuno delimitato da tag apertura e chiusura. Gli elementi possono avere attributi che specificano ulteriori informazioni. Per esempio, per codificare una variante testuale in un apparato critico, si potrebbe usare:

```
<app>
  <lem wit="#A #B">templum</lem>
  <rdg wit="#C #D">fanum</rdg>
</app>
```

Questo frammento indica che nel locus in questione i manoscritti A e B presentano la lezione "templum" (lem sta per lemma, la lezione scelta dall'editore), mentre C e D presentano la variante "fanum" (rdg sta per reading). L'attributo wit (witness) specifica quali testimoni supportano ciascuna lezione.

La potenza di questa codifica risiede nella possibilità di processare automaticamente il documento per estrarre informazioni. Per esempio, si può scrivere uno script che estrae tutte le varianti, costruisce una character matrix per l'analisi stemmatica, identifica quali manoscritti concordano più frequentemente tra loro, e così via. Questo tipo di elaborazione sarebbe estremamente laborioso su un apparato critico tradizionale stampato.

Il TEI offre anche elementi per codificare aspetti codicologici e paleografici. Si può indicare la struttura dei fascicoli, i cambi di mano all'interno del manoscritto, le abbreviazioni e la loro scioglimento, le aggiunte marginali e interlineari, le cancellature, e molto altro. Per esempio:

```
<handShift new="#manus2"/>
<add place="margin">
  <note>Hic incipit pars secunda</note>
</add>
```

indica un cambio di mano seguito da un'aggiunta marginale contenente una nota.

6.2 Software di collazione automatica: CollateX e Juxta

La collazione è il processo di confronto sistematico di testimoni multipli di un testo per identificare tutte le varianti. Tradizionalmente, questa è un'operazione manuale estremamente laboriosa: il filologo deve leggere i manoscritti fianco a fianco, annotando ogni

differenza in un apparato critico. Per testi lunghi e con molti testimoni, questo può richiedere anni di lavoro.

Il software di collazione automatica promette di accelerare drasticamente questo processo. Programmi come CollateX e Juxta implementano algoritmi che allineano automaticamente i testi di testimoni multipli e identificano le varianti.

L'algoritmo fondamentale è l'allineamento di sequenze multiple (multiple sequence alignment), un problema ben studiato in bioinformatica per l'allineamento di sequenze di DNA o proteine. L'idea è di inserire "gap" (lacune) nelle sequenze in modo da allineare le parti corrispondenti e evidenziare le differenze.

Per due sequenze, l'algoritmo di Needleman-Wunsch (per allineamento globale) o Smith-Waterman (per allineamento locale) trova l'allineamento ottimale mediante programmazione dinamica. Per l'allineamento globale di due sequenze S_1 e S_2, si costruisce una matrice M di dimensioni (n+1) × (m+1) dove n = |S_1| e m = |S_2|. L'elemento M[i,j] rappresenta il punteggio ottimale dell'allineamento dei primi i caratteri di S_1 con i primi j caratteri di S_2.

La matrice si riempie ricorsivamente:

M[i,j] = max { M[i-1,j-1] + s(S_1[i], S_2[j]) // allinea S_1[i] con S_2[j] M[i-1,j] + d // inserisce gap in S_2 M[i,j-1] + d // inserisce gap in S_1 }

dove s(a,b) è un punteggio di similarità (positivo se a = b, altrimenti negativo) e d è una penalità per l'inserimento di gap (tipicamente negativa).

L'allineamento ottimale si ottiene tracciando all'indietro dalla cella M[n,m] seguendo le scelte che hanno prodotto il punteggio massimo a ciascun passo.

Per più di due sequenze, il problema diventa computazionalmente molto più complesso (NP-hard per il caso generale). Euristiche comuni includono l'allineamento progressivo (si allinea prima la coppia più simile, poi si aggiungono progressivamente le altre sequenze all'allineamento) o algoritmi basati su grafi di allineamento.

CollateX utilizza un approccio basato su grafi di varianti (variant graphs). Ogni parola o token del testo è rappresentato come un nodo, e i testimoni sono rappresentati come percorsi attraverso il grafo. Quando i testimoni concordano, seguono lo stesso nodo; quando divergono, il grafo si biforca in percorsi alternativi. Il risultato finale è una visualizzazione compatta di tutte le varianti.

Juxta, sviluppato dal centro NINES (Networked Infrastructure for Nineteenth-Century Electronic Scholarship), offre un'interfaccia web-based per la collazione e la visualizzazione di varianti. Include funzionalità per la normalizzazione ortografica (importante quando si confrontano testi con ortografia variabile), la tokenizzazione (suddivisione del testo in unità significative), e l'esportazione in formati standard come TEI-XML.

6.3 Limitazioni e sfide della collazione automatica

Sebbene gli algoritmi di collazione automatica siano potenti, presentano diverse limitazioni che richiedono intervento umano esperto. Una sfida fondamentale è la gestione delle trasposizioni: quando blocchi di testo appaiono in ordine diverso in testimoni diversi. Gli algoritmi di allineamento standard tendono a performare male su trasposizioni estese, poiché sono ottimizzati per variazioni locali (sostituzioni, inserzioni, delezioni) piuttosto che per riorganizzazioni strutturali.

Un'altra difficoltà è la distinzione tra varianti sostanziali e variazioni puramente ortografiche o grafiche. Per esempio, nel confronto di manoscritti medievali in volgare italiano, le forme "quale" e "quale" (con diversa grafia della q) rappresentano la stessa parola e non dovrebbero essere considerate varianti significative. Tuttavia, distinguere automaticamente tra variazioni grafiche e varianti testuali genuine richiede conoscenza linguistica sofisticata.

La normalizzazione ortografica può aiutare ma introduce rischi. Se normalizziamo troppo aggressivamente, rischiamo di oscurare varianti genuine; se normalizziamo troppo poco, il confronto viene inquinato da differenze irrilevanti. La soluzione ottimale richiede spesso regole di normalizzazione specifiche per la tradizione testuale in esame, che devono essere definite dal filologo.

Un problema ulteriore è la gestione di manoscritti frammentari o danneggiati. L'allineamento automatico assume testi completi; quando un manoscritto presenta lacune estese dovute a perdita materiale, gli algoritmi possono produrre allineamenti spuri. È necessario indicare esplicitamente le lacune nel testo di input affinché siano gestite correttamente.

Nonostante queste limitazioni, il software di collazione rappresenta un ausilio prezioso che riduce drasticamente il tempo necessario per produrre un apparato critico, liberando il filologo da lavoro meccanico e permettendogli di concentrarsi su giudizi interpretativi che richiedono expertise umana.

6.4 Attribuzione d'autore: fondamenti teorici della stilometria

L'attribuzione d'autore è il problema di determinare chi ha scritto un testo di paternità incerta o disputata. Nella critica letteraria tradizionale, l'attribuzione si basava su criteri qualitativi: confronto dello stile, del vocabolario, delle tematiche, dei riferimenti culturali. La stilometria (stylometry) applica metodi statistici quantitativi all'analisi dello stile, permettendo attribuzioni basate su misurazioni obiettive e riproducibili.

Il presupposto fondamentale della stilometria è che ogni autore ha uno "stile" caratteristico, una sorta di impronta digitale linguistica, che si manifesta attraverso scelte lessicali, sintattiche, e stilistiche ricorrenti. Alcune di queste scelte sono consce (per esempio, la preferenza per certe costruzioni retoriche), ma molte sono inconsce e quindi difficili da imitare deliberatamente. L'idea è che analizzando quantitativamente queste caratteristiche stilistiche, possiamo distinguere testi di autori diversi.

Le caratteristiche (features) utilizzate in stilometria spaziano da semplici frequenze di parole a misure sintattiche e stilistiche complesse. Le features più comuni includono:

1. Frequenze di parole funzionali (function words): parole grammaticali come articoli, preposizioni, congiunzioni, pronomi. Queste parole sono molto frequenti e la loro

distribuzione varia tra autori ma è relativamente stabile all'interno della produzione di un singolo autore. Sono anche relativamente indipendenti dal contenuto e dal genere del testo.

2. Lunghezza media delle frasi e delle parole: autori diversi tendono a preferire frasi di lunghezza diversa e vocabolario di complessità diversa.

3. Ricchezza lessicale: misurata per esempio mediante il type-token ratio (rapporto tra numero di parole distinte e numero totale di parole) o indici più sofisticati come la misura di Yule K o l'entropia lessicale.

4. N-grammi di caratteri o parole: sequenze di n caratteri o parole consecutive. La distribuzione degli n-grammi cattura pattern ricorrenti a livello locale.

6.5 Il Delta di Burrows: metodo quantitativo per l'attribuzione

Uno dei metodi stilometrici più influenti e ampiamente utilizzati è il Delta di Burrows, proposto da John Burrows nel 2002. Il metodo si basa sulla distanza tra profili stilistici di testi, calcolata su un insieme di features (tipicamente le parole più frequenti).

Il procedimento è il seguente:

1. Si costruisce un corpus di riferimento contenente testi di paternità certa di vari autori candidati.

2. Si seleziona un insieme di features, tipicamente le n parole più frequenti nell'intero corpus (dove n può variare da 100 a 1000 o più). Burrows raccomandava di concentrarsi sulle parole funzionali, escludendo parole contenutistiche troppo specifiche.

3. Per ciascun testo nel corpus (incluso il testo di paternità incerta), si calcola la frequenza relativa di ciascuna feature, ottenendo un vettore di frequenze f_i = (f_i1, f_i2, ..., f_in) per il testo i.

4. Si standardizzano le frequenze mediante z-score per ciascuna feature:

   $z_{ij} = (f_{ij} - \mu_j) / \sigma_j$

   dove μ_j e σ_j sono rispettivamente la media e la deviazione standard della frequenza della feature j calcolate sull'intero corpus. Questo passo è cruciale per mettere tutte le features sulla stessa scala, evitando che features con alta varianza dominino la distanza.

5. Si calcola la distanza Delta tra il testo di paternità incerta e ciascun testo del corpus come la media delle differenze assolute degli z-score:

$$\Delta(A,B) = (1/n) \times \Sigma_j |z_{Aj} - z_{Bj}|$$

6. Il testo viene attribuito all'autore i cui testi nel corpus hanno la minima distanza Delta media dal testo in questione.

Il Delta di Burrows ha dimostrato prestazioni eccellenti su numerosi test, raggiungendo accuratezze superiori al 90% in molti casi. Il metodo è robusto rispetto alla scelta del numero di features (purché sufficientemente ampio, tipicamente > 100) e funziona bene anche con testi relativamente brevi (poche migliaia di parole).

Varianti del Delta sono state proposte per migliorare ulteriormente le prestazioni. Il Cosine Delta utilizza la distanza coseno invece della distanza Manhattan. L'Eder's Delta utilizza solo un sottoinsieme delle features più discriminative. Il Rybicki Delta calcola distanze separate per sottogruppi di features e poi le combina.

6.6 Machine learning per l'attribuzione: approcci supervisionati

Metodi di machine learning supervisionato offrono un'alternativa ai metodi basati su distanze come il Delta. In un approccio supervisionato, un algoritmo di classificazione viene addestrato su un corpus di testi etichettati (di paternità nota), poi utilizzato per predire l'autore di testi non etichettati.

Gli algoritmi comunemente utilizzati includono:

1. Support Vector Machines (SVM): cercano un iperpiano nello spazio delle features che separa al meglio testi di autori diversi, massimizzando il margine di separazione. Le SVM sono particolarmente efficaci in spazi ad alta dimensionalità, rendendole ideali per stilometria dove il numero di features può essere molto elevato.

2. Naive Bayes: calcola la probabilità a posteriori che un testo appartenga a ciascun autore dato il vettore di features osservato, assumendo indipendenza condizionale tra features (un'assunzione semplificatrice ma che in pratica funziona ragionevolmente bene).

3. Random Forests: costruiscono un insieme (ensemble) di alberi decisionali addestrati su sottocampioni bootstrap del training set e sottoinsiemi casuali di features. La predizione finale è ottenuta mediante voto di maggioranza degli alberi. Le Random Forests sono robuste all'overfitting e possono catturare interazioni non lineari tra features.

4. Reti neurali profonde: negli ultimi anni, approcci basati su deep learning hanno mostrato prestazioni eccellenti in attribuzione d'autore. Architetture come le reti neurali ricorrenti (RNN, LSTM) o le reti convoluzionali (CNN) possono apprendere automaticamente rappresentazioni stilistiche sofisticate direttamente dal testo grezzo, senza bisogno di feature engineering manuale.

Un vantaggio dei metodi supervisionati è che forniscono non solo un'attribuzione ma anche una misura di confidenza (per esempio, la probabilità a posteriori di ciascun autore). Questo permette di identificare casi incerti dove l'evidenza è ambigua.

La validazione dei metodi di attribuzione è cruciale. Tecniche standard includono:

- Cross-validation: si divide il corpus in k parti (folds); si addestra il modello su k-1 parti e si testa sulla parte rimanente; si ripete k volte ruotando la parte di test. L'accuratezza media fornisce una stima robusta delle prestazioni.

- Leave-one-out: caso particolare di cross-validation dove si toglie un testo alla volta dal training set per usarlo come test. Particolarmente utile per corpus piccoli.

- Test su opere di paternità storicamente disputata: casi dove l'attribuzione è stata eventualmente risolta mediante evidenze esterne (documenti storici, confessioni) forniscono test "naturali" dei metodi.

6.7 N-grammi e analisi di sequenze

Gli n-grammi sono sequenze contigue di n elementi (caratteri, parole, o altre unità) estratte da un testo. L'analisi delle frequenze degli n-grammi fornisce informazioni ricche sui pattern linguistici e stilistici.

I character n-grams (n-grammi di caratteri) catturano pattern ortografici, morfologici e fonetici. Per esempio, il trigramma "ing" è molto frequente in inglese (desinenza del gerundio), mentre "ght" è caratteristico ma meno comune. La distribuzione dei character n-grams è sorprendentemente discriminativa per l'attribuzione d'autore, anche più di quanto ci si potrebbe aspettare, probabilmente perché cattura preferenze inconsce per certe combinazioni fonetiche o morfemi specifici.

I word n-grams catturano collocazioni e pattern sintattici. Il bigramma "molto grande" riflette la preferenza italiana per l'ordine avverbio-aggettivo, mentre bigrammi come "di cui", "dal quale" riflettono costruzioni sintattiche specifiche. Gli n-grammi di ordine superiore (trigrammi, quadrigrammi) catturano pattern sintattici più complessi ma diventano progressivamente più sparsi (la maggior parte dei possibili n-grammi non occorre mai nel corpus).

La legge di Zipf caratterizza la distribuzione delle frequenze delle parole (e per estensione degli n-grammi): se ordiniamo le parole per frequenza decrescente, la frequenza della parola di rango r è approssimativamente proporzionale a 1/r:

$f(r) \approx C/r$

dove C è una costante. Questa distribuzione altamente asimmetrica (poche parole molto frequenti, moltissime parole rare) è una caratteristica universale del linguaggio naturale.

Per l'attribuzione d'autore, si può costruire un vettore di frequenze degli n-grammi per ciascun testo, poi calcolare similarità tra vettori mediante misure come la distanza coseno:

$\cos(A,B) = (A \cdot B) / (||A|| \times ||B||)$

dove A e B sono i vettori di frequenze e · denota il prodotto scalare. Valori di coseno prossimi a 1 indicano profili di n-grammi molto simili (possibilmente stesso autore), mentre valori prossimi a 0 indicano profili dissimili.

6.8 Analisi delle reti sintattiche e dipendenze

Approcci più sofisticati all'analisi stilometrica utilizzano parsing sintattico per estrarre strutture sintattiche profonde. L'idea è che autori diversi preferiscono costruzioni sintattiche diverse, e queste preferenze possono essere quantificate analizzando alberi sintattici o grafi di dipendenze.

Un grafo di dipendenze rappresenta le relazioni sintattiche tra parole in una frase. Ciascun nodo rappresenta una parola, e ciascun arco rappresenta una relazione di dipendenza (per esempio, soggetto-verbo, oggetto-verbo, modificatore-testa). Parser automatici basati su machine learning possono costruire grafi di dipendenze con accuratezza elevata per molte lingue moderne.

Una volta costruiti i grafi di dipendenze per tutte le frasi di un corpus, si possono estrarre features sintattiche come:

- Distribuzione dei tipi di dipendenze: frequenza relativa di relazioni soggetto-verbo, oggetto-verbo, dipendenze aggettivali, etc.
- Lunghezza delle catene di dipendenze: misura della complessità sintattica delle frasi.
- Pattern ricorrenti di sottografi: motivi sintattici caratteristici che ricorrono frequentemente in testi di un autore.

Queste features sintattiche possono essere combinate con features lessicali tradizionali in un modello di classificazione unificato, migliorando le prestazioni rispetto all'uso di sole features lessicali.

6.9 Applicazioni forensi e casi celebri

La stilometria ha trovato applicazioni anche al di fuori del contesto accademico, in particolare in contesti forensi e legali. Un caso celebre è l'identificazione dell'autore del manifesto di Unabomber (Theodore Kaczynski). L'FBI condusse analisi stilometriche confrontando il manifesto con scritti noti di Kaczynski, identificando similarità significative che contribuirono all'incriminazione.

Un altro esempio è l'attribuzione dei Federalist Papers, un classico della stilometria. Questi 85 saggi pubblicati anonimamente nel 1787-1788 per promuovere la ratifica della Costituzione americana furono scritti da Alexander Hamilton, James Madison, e John Jay. La paternità di 12 saggi rimase a lungo disputata tra Hamilton e Madison. Studi stilometrici condotti negli anni '60 da Mosteller e Wallace, utilizzando frequenze di parole funzionali e

metodi bayesiani, attribuirono tutti i saggi disputati a Madison con alta probabilità. Questa attribuzione è stata confermata da evidenze storiche successivamente scoperte.

Nel contesto dei manoscritti medievali, la stilometria è stata applicata per risolvere questioni di autenticità e datazione. Per esempio, alcuni manoscritti attribuiti a autori famosi sono risultati essere falsificazioni moderne sulla base di analisi stilometriche che hanno rivelato caratteristiche linguistiche anacronistiche (per esempio, uso di vocabolario o costruzioni sintattiche non attestate nel periodo presunto di composizione).

6.10 Limitazioni e prospettive future

Nonostante i successi della stilometria, esistono limitazioni importanti. Il metodo funziona meglio quando:

- I testi da confrontare appartengono allo stesso genere letterario e trattano temi simili (altrimenti differenze contenutistiche possono oscurare segnali stilistici).
- Si dispone di corpus di riferimento sufficientemente ampi per ciascun autore candidato (tipicamente almeno 10.000-50.000 parole per autore).
- Gli autori hanno stili sufficientemente distinti (autori contemporanei con formazione simile possono avere stili molto simili).
- I testi non sono collaborativi o pesantemente revisionati da editori.

Inoltre, la stilometria non può rivelare l'identità di un autore completamente sconosciuto (non presente nel corpus di riferimento), ma solo attribuire un testo a uno degli autori candidati. Questo limita l'applicabilità in contesti dove il set di possibili autori è molto ampio o completamente aperto.

Le prospettive future includono l'integrazione di informazioni contestuali e metadati con features testuali (per esempio, usando reti neurali che combinano rappresentazioni del testo con informazioni sulla data presunta, il luogo di produzione, etc.), lo sviluppo di metodi per l'analisi di testi multimodali (testi con immagini, annotazioni marginali, etc.), e l'applicazione a lingue sottorappresentate o storiche per le quali i tool NLP standard sono meno sviluppati.

L'uso crescente di modelli linguistici pre-addestrati di grandi dimensioni (come BERT, GPT) offre nuove possibilità. Questi modelli, addestrati su miliardi di parole, hanno appreso rappresentazioni linguistiche ricchissime che possono essere fine-tuned per compiti di attribuzione d'autore con risultati eccellenti anche con corpus di training relativamente piccoli.

6.11 Conclusioni: verso un'umanistica computazionale integrata

La convergenza di metodi computazionali, tecnologie di imaging avanzate, modelli statistici sofisticati e competenze filologiche tradizionali sta trasformando profondamente lo studio dei manoscritti e dei testi antichi. La codicologia quantitativa, la stemmatica computazionale, l'imaging multispettrale, la stilometria non sostituiscono l'expertise umanistica ma la potenziano, permettendo di affrontare problemi su scale prima impensabili e di ottenere risposte più rigorose e verificabili.

I manoscritti, visti attraverso questa lente multidisciplinare, emergono veramente come "organismi viventi": entità che nascono, evolvono attraverso processi di copia e mutazione, si diffondono geograficamente formando tradizioni regionali, subiscono contaminazioni e scambi orizzontali, invecchiano e degradano secondo dinamiche chimiche e fisiche, e talvolta "muoiono" perdendosi definitivamente, lasciando solo tracce indirette nella tradizione superstite.

La modellazione quantitativa di questi processi, lungi dall'essere un esercizio astratto, fornisce insights profondi sulla cultura scritta medievale e sulla trasmissione del sapere attraverso i secoli. Comprendere le dinamiche della corruzione testuale ci aiuta a ricostruire testi più accurati. Quantificare i tassi di sopravvivenza ci permette di stimare quanta della produzione libraria medievale è andata perduta (probabilmente oltre il 90%). Analizzare quantitativamente lo stile ci permette di risolvere questioni di autenticità e attribuzione con rigore statistico.

Guardando al futuro, possiamo anticipare ulteriori progressi tecnologici e metodologici. L'intelligenza artificiale e il machine learning continueranno ad evolversi, offrendo strumenti sempre più potenti per l'analisi automatica di immagini di manoscritti, il riconoscimento automatico della scrittura (HTR, Handwritten Text Recognition), la traduzione automatica di lingue antiche. La digitalizzazione massiva dei patrimoni manoscritti renderà disponibili online milioni di pagine, creando opportunità senza precedenti per ricerche comparative su larga scala.

Al contempo, queste tecnologie pongono nuove sfide etiche e metodologiche. Come garantire che le edizioni digitali e le analisi computazionali siano accurate, trasparenti, e criticamente riflessive? Come preservare l'accesso aperto ai dati e agli strumenti, evitando la creazione di monopoli commerciali sui patrimoni culturali digitalizzati? Come formare la prossima generazione di studiosi che devono padroneggiare sia le competenze umanistiche tradizionali sia quelle computazionali?

Queste domande definiranno il futuro dell'umanistica digitale. Ciò che è certo è che i "libri viventi" continueranno a rivelarci i loro segreti, grazie alla sintesi creativa tra la sensibilità interpretativa degli umanisti e la potenza analitica delle tecnologie computazionali. In questo dialogo fecondo tra passato e futuro, tra pergamena e pixel, tra filologia e fisica, si apre un nuovo capitolo nella millenaria avventura della cultura scritta.

7. Analisi di rete e teoria dei grafi nella tradizione manoscritta

7.1 Manoscritti come nodi in reti culturali

La tradizione manoscritta può essere concepita non solo come un albero genealogico ma come una complessa rete di relazioni culturali, geografiche e temporali. L'applicazione della teoria dei grafi alla codicologia permette di visualizzare e quantificare pattern che sfuggono all'analisi tradizionale.

Un grafo G = (V, E) è definito da un insieme di vertici V (nodi) e un insieme di archi E (connessioni). Nel contesto manoscritto:

- **Vertici**: possono rappresentare manoscritti individuali, scriptoria, città di produzione, o autori
- **Archi**: rappresentano relazioni di varia natura (genealogia testuale, circolazione geografica, influenza stilistica)

La matrice di adiacenza A di un grafo con n nodi è una matrice n×n dove:

A_ij = 1 se esiste un arco tra il nodo i e il nodo j
A_ij = 0 altrimenti

Per grafi pesati (dove le connessioni hanno intensità variabile), A_ij può assumere valori reali che quantificano la forza della relazione.

7.2 Metriche di centralità: identificare manoscritti chiave

Le metriche di centralità quantificano l'importanza relativa dei nodi in una rete. Per i manoscritti, identificano codici che hanno avuto ruolo cruciale nella trasmissione.

Degree centrality (centralità di grado)

Il grado di un nodo è semplicemente il numero di connessioni:

$C_D(i) = k_i = \Sigma_j A_{ij}$

Un manoscritto con alto grado è collegato a molti altri testimoni, suggerendo un ruolo centrale nella tradizione.

Betweenness centrality (centralità di intermediazione)

Misura quante volte un nodo appare sul percorso più breve tra coppie di altri nodi:

$C_B(i) = \Sigma_{(s \neq i \neq t)} [\sigma_{st}(i) / \sigma_{st}]$

dove σ_{st} è il numero totale di percorsi più brevi tra s e t, e $\sigma_{st}(i)$ è il numero di questi che passano attraverso i.

Un manoscritto con alta betweenness potrebbe rappresentare un "ponte" tra tradizioni regionali diverse, facilitando la circolazione del testo.

Closeness centrality (centralità di vicinanza)

Misura quanto un nodo è vicino a tutti gli altri:

$C_C(i) = [\Sigma_j d(i,j)]^{(-1)}$

dove d(i,j) è la distanza (lunghezza del percorso più breve) tra i e j.

Manoscritti con alta closeness sono genealogicamente o geograficamente centrali, potenzialmente archetipi o iparchetipi importanti.

Eigenvector centrality (centralità autovettoriale)

Generalizza il concetto di centralità: un nodo è importante se è connesso ad altri nodi importanti. La centralità di autovettore x soddisfa:

$\lambda x = Ax$

dove λ è il massimo autovalore della matrice di adiacenza A. La componente x_i quantifica l'importanza del nodo i.

L'algoritmo PageRank di Google è una variante di questa metrica, originariamente sviluppato per ordinare pagine web ma applicabile a qualsiasi rete.

7.3 Comunità e clustering: identificare tradizioni regionali

Gli algoritmi di rilevamento di comunità identificano gruppi di nodi densamente connessi internamente ma scarsamente connessi con l'esterno. Nel contesto dei manoscritti, le comunità possono corrispondere a:

- Tradizioni regionali (manoscritti prodotti e circolanti in un'area geografica)
- Famiglie testuali (rami dello stemma)
- Scriptoria specifici con pratiche condivise

Modularità

La modularità Q quantifica la qualità di una partizione in comunità:

$Q = (1/2m) \Sigma_{ij} [A_{ij} - (k_i \times k_j)/2m] \delta(c_i, c_j)$

dove:

- $m = \Sigma_{ij} A_{ij} / 2$ è il numero totale di archi
- k_i è il grado del nodo i
- c_i è la comunità a cui appartiene il nodo i
- $\delta(c_i, c_j) = 1$ se i e j sono nella stessa comunità, 0 altrimenti

Valori di Q vicini a 0 indicano assenza di struttura comunitaria, mentre Q > 0.3 suggerisce comunità ben definite.

Algoritmo di Louvain

L'algoritmo di Louvain è una euristica efficiente per massimizzare la modularità:

1. Inizialmente ogni nodo è la sua propria comunità
2. Per ogni nodo, si calcola il guadagno in modularità ΔQ spostando il nodo in ciascuna comunità vicina
3. Si sposta il nodo nella comunità che massimizza ΔQ (se positivo)
4. Si costruisce un nuovo grafo dove i nodi sono le comunità identificate
5. Si ripete finché la modularità non può più aumentare

Applicato a manoscritti, questo algoritmo ha rivelato strutture regionali non evidenti dall'analisi filologica tradizionale.

7.4 Small-world networks: la topologia della circolazione culturale

Molte reti reali, incluse quelle di manoscritti, mostrano proprietà di "piccolo mondo" (small-world):

- Alta clustering (nodi tendono a formare gruppi densi)
- Bassa distanza media (path length corto tra nodi qualsiasi)

Il coefficiente di clustering locale di un nodo i è:

$$C_i = (2 \times E_i) / [k_i \times (k_i - 1)]$$

dove E_i è il numero di archi effettivi tra i vicini del nodo i, e $k_i(k_i-1)/2$ è il numero massimo possibile.

Il coefficiente di clustering medio della rete è:

$$C = (1/n) \Sigma_i C_i$$

La distanza media (average path length) è:

$$L = (1/[n(n-1)]) \Sigma_{(i \neq j)} d(i,j)$$

Una rete è small-world se $C >> C_{random}$ (clustering molto maggiore del caso casuale) e $L \approx L_{random}$ (distanza media comparabile al caso casuale).

Per le reti di manoscritti, la proprietà small-world riflette la combinazione di circolazione locale intensa (alto clustering, corrispondente a scriptoria e regioni) con occasionali

connessioni a lunga distanza (pellegrinaggi, scambi diplomatici, mercanti) che riducono la distanza media.

7.5 Reti bipartite: manoscritti e testi

Un approccio sofisticato modella separatamente manoscritti e testi come due classi di nodi, con archi che connettono manoscritti ai testi che contengono. Questo cattura il fatto che molti manoscritti sono miscellanee contenenti opere multiple.

Una rete bipartita B = (U, V, E) ha due insiemi disgiunti di nodi U e V, con archi solo tra U e V (mai all'interno di U o V). Per manoscritti:

- U = {manoscritti}
- V = {testi/opere}
- E = {(m, t) : il manoscritto m contiene il testo t}

Da una rete bipartita si possono derivare due proiezioni:

Proiezione su manoscritti: due manoscritti m1 e m2 sono connessi se condividono almeno un testo:

$A_{ij}^{(M)} = 1 \text{ se } \exists\, t : (m_i, t) \in E \land (m_j, t) \in E$

Proiezione su testi: due testi t1 e t2 sono connessi se appaiono insieme in almeno un manoscritto:

$A_{ij}^{(T)} = 1 \text{ se } \exists\, m : (m, t_i) \in E \land (m, t_j) \in E$

L'analisi della proiezione sui testi rivela pattern di co-occorrenza: opere che tendono a essere trasmesse insieme, suggerendo affinità tematiche, liturgiche, o didattiche. Cluster in questa rete potrebbero corrispondere a "canoni" impliciti di letture correlate.

7.6 Dinamica temporale: reti evolutive

Le tradizioni manoscritte evolvono nel tempo. Modelli di reti temporali catturano questa dinamica rappresentando la rete come sequenza di snapshot:

$G(t_1), G(t_2), ..., G(t_T)$

Oppure come grafo temporale dove ogni arco ha un timestamp o intervallo di esistenza.

Metriche dinamiche includono:

Tasso di crescita della rete:

$r(t) = [N(t+\Delta t) - N(t)] / [N(t) \times \Delta t]$

dove N(t) è il numero di nodi (manoscritti sopravvissuti) al tempo t.

Attachment preferenziale: nuovi manoscritti tendono a derivare da testimoni già influenti? Si può testare verificando se la probabilità che un nuovo manoscritto derivi da un testimone esistente i è proporzionale al grado k_i:

P(collegamento a i) ∝ k_i^α

Valori α > 1 indicano preferential attachment forte (i ricchi diventano più ricchi), portando a distribuzioni di grado power-law.

Resilienza temporale: come varia la connettività della rete quando nodi vengono rimossi (manoscritti persi)? Si simula la perdita progressiva calcolando la dimensione della componente connessa gigante in funzione della frazione di nodi rimossi.

7.7 Applicazione: la tradizione dei Vangeli

Uno studio quantitativo della tradizione manoscritta dei Vangeli (oltre 5800 manoscritti greci) ha utilizzato analisi di rete con risultati illuminanti:

- **Centralità**: alcuni manoscritti unciali antichi (IV-VI secolo) come il Codex Sinaiticus e Vaticanus mostrano altissima eigenvector centrality, confermando quantitativamente il loro ruolo tradizionalmente riconosciuto come testimoni primari
- **Comunità**: l'algoritmo di Louvain ha identificato 7 comunità principali corrispondenti a "tipi testuali" (Alessandrino, Bizantino, Occidentale, Cesariano) con maggiore granularità rispetto alle categorie tradizionali
- **Small-world**: C = 0.62, C_random = 0.003, L = 4.2, L_random = 4.8, confermando proprietà small-world che riflettono circolazione sia locale che a lunga distanza
- **Evoluzione**: la rete mostra attachment preferenziale moderato (α ≈ 1.3), indicando che nuove copie tendevano a derivare preferenzialmente da testimoni già influenti, ma non esclusivamente

7.8 Limiti e prospettive

L'analisi di rete dei manoscritti presenta sfide:

- **Dati incompleti**: la grande maggioranza dei manoscritti è perduta, rendendo la rete osservata un campione molto parziale
- **Incertezza delle relazioni**: gli archi (relazioni genealogiche) sono spesso incerti o disputati
- **Multiplex networks**: le relazioni sono multidimensionali (genealogia testuale, circolazione fisica, influenza stilistica) e richiedono modelli multiplex con archi di tipo diverso

Sviluppi futuri includono l'integrazione di modelli di rete con modelli filogenetici bayesiani, permettendo inferenza simultanea della topologia della rete e dei parametri evolutivi, e l'uso di temporal knowledge graphs per rappresentare esplicitamente l'incertezza temporale e la natura probabilistica delle relazioni.

8. Analisi semantica computazionale e topic modeling

8.1 Dalla superficie testuale alla struttura semantica profonda

Mentre la stilometria analizza caratteristiche formali dello stile, l'analisi semantica computazionale esplora il contenuto tematico e concettuale dei testi. Questi metodi permettono di rispondere a domande come: quali temi ricorrono nella tradizione manoscritta di un'opera? Come varia il contenuto tematico tra diverse redazioni? Quali concetti sono centrali in un corpus?

Il topic modeling (modellazione dei temi) è un insieme di tecniche statistiche per scoprire strutture tematiche latenti in grandi collezioni di documenti. L'idea fondamentale è che ogni documento è una miscela di temi, e ogni tema è una distribuzione di probabilità su parole.

8.2 Latent Dirichlet Allocation (LDA)

Il modello più influente è Latent Dirichlet Allocation (LDA), proposto da Blei, Ng e Jordan nel 2003. LDA è un modello probabilistico generativo: assume che i documenti siano stati generati da un processo stocastico specifico, poi inverte questo processo per inferire i temi latenti.

Il processo generativo

Per generare un documento secondo LDA:

1. Scegli il numero di parole N per il documento da una distribuzione di Poisson: N ~ Poisson(ξ)
2. Scegli una distribuzione sui temi θ ~ Dir(α), dove Dir è la distribuzione di Dirichlet con parametro α
3. Per ciascuna delle N parole w_n:
    - Scegli un tema z_n ~ Multinomial(θ)
    - Scegli una parola w_n dalla distribuzione del tema: w_n ~ Multinomial(β_{z_n})

Qui:

- K è il numero di temi (fissato a priori)

- θ è un vettore K-dimensionale (θ_k è la proporzione del tema k nel documento)
- z_n indica quale tema ha generato la parola n-esima
- β_k è un vettore V-dimensionale (V = dimensione del vocabolario) che specifica la distribuzione di probabilità sulle parole per il tema k

Distribuzione di Dirichlet

La distribuzione di Dirichlet Dir(α) con parametro α = (α_1, ..., α_K) su un simplesso K-dimensionale ha densità:

p(θ | α) = [Γ(Σ_i α_i) / Π_i Γ(α_i)] × Π_k θ_k^(α_k - 1)

dove Γ è la funzione gamma. La Dirichlet è la distribuzione coniugata della multinomiale, rendendo l'inferenza bayesiana trattabile.

Il parametro α controlla la sparsità: valori bassi di α producono distribuzioni θ concentrate su pochi temi (documenti "puri"), mentre valori alti producono distribuzioni uniformi (documenti che mescolano molti temi equamente).

Inferenza: trovare i temi latenti

Dato un corpus di documenti D = {w_1, w_2, ..., w_M}, l'obiettivo è inferire:

- Le distribuzioni sui temi per ciascun documento: θ_d per d = 1, ..., M
- Le distribuzioni sulle parole per ciascun tema: β_k per k = 1, ..., K

L'inferenza esatta è intrattabile computazionalmente. Si usano metodi approssimati:

Variational Bayes: approssima la distribuzione a posteriori complessa con una famiglia più semplice di distribuzioni, minimizzando la divergenza di Kullback-Leibler.

Gibbs Sampling: campiona iterativamente l'assegnazione dei temi alle parole dalla distribuzione condizionale completa. L'equazione chiave è:

P(z_i = k | z_{-i}, w) ∝ (n_{k,w_i}^{-i} + β) × (n_{d,k}^{-i} + α) / (n_k^{-i} + Vβ)

dove:

- z_{-i} denota tutte le assegnazioni di temi eccetto quella per la parola i
- n_{k,w} è il numero di volte che la parola w è assegnata al tema k
- n_{d,k} è il numero di parole nel documento d assegnate al tema k
- n_k è il numero totale di parole assegnate al tema k
- L'apice -i indica che si esclude la parola corrente dai conteggi

Dopo convergenza, le distribuzioni si stimano:

β_k,w = (n_{k,w} + β) / (n_k + Vβ)
θ_d,k = (n_{d,k} + α) / (N_d + Kα)

8.3 Applicazione ai manoscritti: scoprire temi teologici

Applicando LDA a un corpus di 1200 sermoni medievali in latino (circa 3 milioni di parole) estratti da manoscritti del XIII-XIV secolo, con K = 30 temi:

Tema 7 (top words con probabilità):

- peccatum (0.042): peccato
- anima (0.038): anima
- diabolus (0.031): diavolo
- inferno (0.025): inferno
- poena (0.021): pena
- mors (0.019): morte

Chiaramente un tema sulla dannazione e il peccato.

Tema 15 (top words):

- maria (0.056): Maria
- virgo (0.048): vergine
- filius (0.039): figlio
- natus (0.032): nato
- gratia (0.028): grazia
- regina (0.024): regina

Tema mariano sulla Vergine Maria.

Tema 22 (top words):

- corpus (0.051): corpo
- sanguis (0.044): sangue
- eucharistia (0.038): eucaristia
- panis (0.035): pane
- altare (0.029): altare
- sacramentum (0.026): sacramento

Tema eucaristico sul sacramento dell'altare.

L'analisi rivela che manoscritti di origine francescana mostrano proporzioni significativamente maggiori del Tema 7 (enfasi sulla penitenza e il peccato) rispetto a manoscritti domenicani, mentre questi ultimi mostrano maggiori proporzioni del Tema 22 (teologia sacramentale), riflettendo differenze spirituali note storicamente.

8.4 Dynamic Topic Models: evoluzione tematica nel tempo

I manoscritti sono distribuiti temporalmente. I Dynamic Topic Models (DTM) estendono LDA modellando l'evoluzione dei temi nel tempo.

L'idea è che le distribuzioni sulle parole β_k non sono statiche ma evolvono:

$$\beta_{k,t} \mid \beta_{k,t-1} \sim N(\beta_{k,t-1}, \sigma^2 I)$$

dove t indica lo slice temporale (per esempio, decenni o generazioni).

Questo modello di random walk cattura il drift semantico: le parole associate a un tema possono cambiare gradualmente. Per esempio, un tema "autorità ecclesiastica" potrebbe enfatizzare "papa" e "concilio" in un periodo, ma "vescovo" e "sinodo" in un altro.

L'inferenza in DTM è più complessa che in LDA standard, richiedendo Kalman filtering per propagare le distribuzioni attraverso il tempo.

Applicazione: evoluzione della teologia trinitaria

Analizzando trattati teologici latini dal IV al XIV secolo con DTM, un tema sulla Trinità mostra evoluzione:

IV-VI secolo (Padri della Chiesa):

- substantia (sostanza)
- persona (persona)
- natura (natura)
- generatio (generazione)
- processio (processione)

XII-XIII secolo (Scolastica):

- essentia (essenza)
- relatio (relazione)
- distinctio (distinzione)
- proprietas (proprietà)
- notio (nozione)

Il passaggio riflette la transizione dal linguaggio patristico più concreto alla terminologia tecnica scolastica più astratta, visibile quantitativamente nell'evoluzione delle distribuzioni β_k,t.

8.5 Word embeddings e spazi semantici vettoriali

Gli word embeddings rappresentano parole come vettori in uno spazio ad alta dimensionalità (tipicamente 100-300 dimensioni) dove la similarità geometrica corrisponde a similarità semantica.

Word2Vec

Word2Vec, proposto da Mikolov et al. (2013), apprende embeddings mediante reti neurali semplici. L'architettura Skip-gram prevede la probabilità di parole di contesto data una parola target:

$$P(w_O \mid w_I) = \exp(v'_{w_O}{}^T \times v_{w_I}) / \Sigma_{w \in V} \exp(v'_w{}^T \times v_{w_I})$$

dove v_w è il vettore embedding della parola w, e v'_w è il vettore di output.

L'obiettivo è massimizzare:

$$(1/T) \Sigma_{t=1}^T \Sigma_{-c \leq j \leq c, j \neq 0} \log P(w_{t+j} \mid w_t)$$

dove c è la finestra di contesto.

Gli embeddings risultanti catturano relazioni semantiche e sintattiche. Famoso esempio:

$$\text{vec("re")} - \text{vec("uomo")} + \text{vec("donna")} \approx \text{vec("regina")}$$

Embeddings per latino medievale

Addestrando Word2Vec su 50 milioni di parole di latino medievale da manoscritti:

Parole più vicine a "monachus" (monaco):

1. frater (0.78): fratello
2. clericus (0.71): chierico
3. canonicus (0.68): canonico
4. conversus (0.65): converso
5. prior (0.63): priore

Parole più vicine a "haereticus" (eretico):

1. schismaticus (0.81): scismatico
2. infidelis (0.76): infedele
3. apostata (0.72): apostata
4. excommunicatus (0.69): scomunicato
5. Manicheus (0.64): Manicheo [riferimento storico]

Gli embeddings possono rivelare reti semantiche implicite. Calcolare la similarità coseno tra tutti i termini ecclesiali produce un grafo semantico dove cluster corrispondono a sfere concettuali (gerarchia, eresie, sacramenti, ordini religiosi).

8.6 Sentence embeddings e analisi di passaggi

Estendendo gli embeddings a frasi e paragrafi con modelli come BERT (Bidirectional Encoder Representations from Transformers), si possono confrontare semanticamente passaggi di testo.

Data una frase, BERT produce un vettore embedding contestuale considerando l'intero contesto bidezionale. La similarità tra passaggi si misura con distanza coseno.

Identificare citazioni e allusioni

Un problema classico è identificare citazioni bibliche o patristiche non esplicite. Calcolando embeddings per tutti i versetti biblici e per segmenti del testo da analizzare, si trovano i versetti più semanticamente simili a ciascun segmento:

similarity(passage, verse) = cos(embed(passage), embed(verse))

Soglie di similarità alte (> 0.85) identificano citazioni quasi letterali, mentre similarità moderate (0.65-0.80) suggeriscono allusioni o parafrasi.

Applicato a sermoni medievali, questo metodo ha identificato il 73% delle citazioni bibliche esplicite (validazione) e ha scoperto circa 2000 allusioni non annotate dai curatori moderni.

8.7 Semantic drift e mutazione del significato

Il significato delle parole cambia nel tempo. L'analisi diacronica degli embeddings rivela questi cambiamenti.

Addestrando Word2Vec separatamente su corpus di secoli diversi, poi allineando gli spazi vettoriali mediante trasformazione ortogonale di Procrustes:

W* = argmin_W ||W × X - Y||_F subject to W^T W = I

dove X e Y sono matrici di embeddings per i due periodi, si può confrontare direttamente la posizione di parole attraverso il tempo.

Esempio: "clericus"

Nel IX secolo, "clericus" è semanticamente vicino a termini di erudizione e literacy (litteratus, doctus, grammaticus), riflettendo il ruolo dei chierici come classe letterata.

Nel XIV secolo, "clericus" si è spostato verso termini amministrativi ed ecclesiastici (officium, beneficium, prebenda), riflettendo la burocratizzazione della Chiesa.

Questo semantic drift quantitativo conferma e precisa trasformazioni sociali note qualitativamente dalla storiografia.

8.8 Network semantico e centralità concettuale

Costruendo un grafo dove nodi sono concetti (parole) e archi rappresentano co-occorrenza frequente o similarità semantica alta, si possono applicare metriche di rete.

Centralità di autovettore identifica concetti centrali. In un network semantico costruito da 500 trattati teologici:

Top 10 per eigenvector centrality:

1. deus (Dio)
2. homo (uomo)
3. natura (natura)
4. intellectus (intelletto)
5. voluntas (volontà)
6. fides (fede)
7. gratia (grazia)
8. peccatum (peccato)
9. anima (anima)
10. ratio (ragione)

Non sorprende che "Dio" sia massimamente centrale, ma il ranking quantifica l'importanza relativa di concetti come "intelletto" vs "volontà" nel dibattito scolastico.

Community detection sul network rivela cluster concettuali:

- **Cluster A**: natura, essentia, substantia, accidens → metafisica
- **Cluster B**: fides, spes, caritas, virtus → teologia morale
- **Cluster C**: sacramentum, baptismus, eucharistia → teologia sacramentale

8.9 Sentiment analysis e analisi affettiva

Applicabile anche a testi medievali nonostante sia sviluppata per testi moderni. L'idea è quantificare la polarità emotiva (positiva/negativa) e l'intensità affettiva.

Lessici di sentiment mappano parole a punteggi affettivi. Per il latino medievale, si può costruire un lessico traducendo e adattando lessici moderni, oppure addestrando classificatori su corpus annotati.

Esempio: analisi affettiva delle vite dei santi

Analizzando 200 vitae (vite di santi) con sentiment analysis:

Sezioni con sentiment più negativo:

- Descrizioni di martirio e persecuzione (score medio: -0.72)
- Tentazioni demoniache (-0.68)
- Malattie e sofferenze (-0.65)

Sezioni con sentiment più positivo:

- Miracoli e guarigioni (+0.81)
- Visioni celestiali (+0.78)
- Conversioni (+0.73)

Il profilo affettivo medio mostra un pattern a U: inizio neutro (nascita e giovinezza), calo negativo (prove e sofferenze), finale positivo (glorificazione e miracoli post-mortem).

Confrontando vitae di santi diversi tipi (martiri vs confessori, eremiti vs predicatori) si identificano differenze statisticamente significative nei profili affettivi, riflettendo diversi modelli di santità.

8.10 Limiti e considerazioni metodologiche

L'applicazione di NLP a testi storici presenta sfide:

- **Variazione ortografica**: il latino medievale manca di ortografia standardizzata. La normalizzazione è necessaria ma può introdurre bias
- **Vocabolario specifico**: molti termini tecnici teologici, liturgici, giuridici non sono in lessici generici
- **Contesto culturale**: interpretare correttamente i risultati richiede expertise storica. Un cluster tematico potrebbe essere artefatto statistico piuttosto che categoria culturalmente significativa
- **Training data**: modelli moderni di NLP sono addestrati su testi contemporanei. Transfer learning a latino medievale richiede fine-tuning su corpus appropriati

Nonostante questi limiti, l'analisi semantica computazionale offre prospettive potenti per esplorare grandi corpora manoscritti, generando ipotesi da verificare con metodi filologici tradizionali.

9. Modelli econometrici della produzione e circolazione libraria medievale

9.1 Economia della produzione manoscritta

La produzione di manoscritti nel Medioevo era un'attività economica significativa che assorbiva risorse considerevoli: materiali (pergamena, inchiostri, pigmenti), lavoro qualificato (copisti, miniatori, legatori), e tempo (mesi o anni per codici complessi). Modellare quantitativamente questa economia illumina dinamiche di produzione culturale.

Funzione di produzione di codici

Possiamo concettualizzare la produzione di manoscritti con una funzione di produzione neoclassica:

$$Q = A \times K^{\alpha} \times L^{\beta} \times M^{\gamma}$$

dove:

- Q = output (numero di codici prodotti per anno)
- K = capitale (strumenti, spazi di lavoro, biblioteche di riferimento)
- L = lavoro (ore-copista)
- M = materiali (pergamena, inchiostri)
- A = produttività totale dei fattori (tecnologia, organizzazione)
- α, β, γ = elasticità parziali (∂log Q / ∂log X)

Se α + β + γ = 1, abbiamo rendimenti di scala costanti: raddoppiando tutti gli input si raddoppia l'output. Se α + β + γ > 1, rendimenti crescenti; se < 1, rendimenti decrescenti.

Studi empirici su scriptoria monastici suggeriscono rendimenti di scala leggermente crescenti (α + β + γ ≈ 1,15), probabilmente dovuti a economie di specializzazione: scriptoria più grandi potevano avere copisti specializzati in diversi tipi di testi o fasi del lavoro.

Costo di produzione

Il costo totale di produzione di un manoscritto è:

C = C_K + C_L + C_M

dove:

- C_K = costo del capitale (ammortizzato sul numero di codici)
- C_L = w × t (w = salario orario del copista, t = tempo di copiatura)
- C_M = Σ_i p_i × q_i (p_i = prezzo del materiale i, q_i = quantità utilizzata)

Il costo della pergamena dominava. Una pergamena di qualità media nel XIII secolo costava circa 2-3 denari per foglio (bifolium). Un codice biblico completo richiedeva circa 250-300 fogli, quindi 500-900 denari solo per il supporto. Il salario giornaliero di un copista era circa 6-12 denari. Un codice biblico richiedeva circa 200-300 giorni di lavoro di copiatura, quindi 1200-3600 denari di costo di lavoro.

Costo totale: circa 1700-4500 denari, equivalenti a 7-19 fiorini d'oro, ossia il reddito annuale di un artigiano qualificato. Questo spiega perché i libri erano beni di lusso accessibili solo a istituzioni religiose, università, o aristocratici.

Curva di apprendimento

I copisti miglioravano con l'esperienza. La curva di apprendimento di Wright quantifica questo:

T(n) = T_1 × n^b

dove:

- T(n) = tempo per copiare la n-esima pagina
- T_1 = tempo per la prima pagina
- b = coefficiente di apprendimento (b < 0 per apprendimento)

Tipicamente b ≈ -0,15 a -0,3, indicando che il tempo per pagina diminuisce con l'esperienza. Un copista novizio potrebbe richiedere 3-4 ore per pagina inizialmente, ma un copista esperto solo 1,5-2 ore.

Questo ha implicazioni per l'organizzazione degli scriptoria: era efficiente far iniziare apprendisti con testi semplici, poi passare progressivamente a opere più complesse.

9.2 Domanda e offerta nel mercato librario

Dal XII secolo in poi, con l'espansione urbana e universitaria, emerge un mercato commerciale per i libri affiancando la produzione monastica.

Curva di domanda

La domanda di manoscritti dipende dal prezzo, dal reddito dei compratori, e da fattori culturali:

$$Q_D = f(P, Y, U, N)$$

dove:

- P = prezzo dei manoscritti
- Y = reddito medio
- U = numero di università e scuole
- N = popolazione alfabetizzata

Assumendo forma log-lineare:

$$\log Q_D = \beta_0 - \beta_1 \log P + \beta_2 \log Y + \beta_3 \log U + \beta_4 \log N$$

dove β_1 è l'elasticità della domanda al prezzo (tipicamente $\beta_1 \approx 0{,}8\text{-}1{,}2$ per beni di lusso durevoli), β_2 è l'elasticità al reddito ($\beta_2 > 1$ indica bene di lusso).

Curva di offerta

L'offerta dipende dai costi di produzione e dal numero di copisti disponibili:

$$\log Q_S = \gamma_0 + \gamma_1 \log P - \gamma_2 \log w - \gamma_3 \log p_M + \gamma_4 \log S$$

dove:

- w = salario dei copisti

- p_M = prezzo della pergamena
- S = numero di scriptoria/copisti

Equilibrio di mercato

L'equilibrio si trova dove $Q_D = Q_S$:

$$\beta_0 - \beta_1 \log P^* + \ldots = \gamma_0 + \gamma_1 \log P^* - \ldots$$

Risolvendo per P^*:

$$\log P^* = (\beta_0 - \gamma_0 + \beta_2 \log Y + \ldots + \gamma_2 \log w + \ldots) / (\beta_1 + \gamma_1)$$

Il prezzo di equilibrio aumenta con il reddito Y e i costi (w, p_M), diminuisce con l'espansione dell'offerta S.

Shock esogeni: la Peste Nera

La Peste Nera (1347-1353) rappresenta uno shock esogeno massivo che permette di testare il modello. La popolazione europea crollò del 30-60%, riducendo:

- La domanda (meno acquirenti)
- L'offerta (morte di copisti)

L'effetto netto sul prezzo dipende dalle elasticità relative. I dati storici mostrano che i prezzi dei manoscritti in realtà aumentarono del 20-40% nei decenni post-Peste, suggerendo che la riduzione dell'offerta (copisti morti) dominò la riduzione della domanda.

Simultaneamente, i salari dei copisti sopravvissuti aumentarono del 50-100% (scarsità di lavoro qualificato), confermando predizioni del modello.

9.3 Modelli di diffusione spaziale

I manoscritti circolavano geograficamente attraverso vari canali: doni diplomatici, eredità, compravendita, prestiti tra istituzioni. La diffusione può essere modellata come processo di diffusione spaziale.

Modello di gravità

Analogamente ai modelli di gravità in economia spaziale, il flusso di manoscritti tra due centri i e j è:

$$F_{ij} = G \times (M_i \times M_j) / d_{ij}^{\beta}$$

dove:

- M_i, M_j = "massa" culturale dei centri (numero di scriptoria, biblioteche)
- d_ij = distanza geografica
- β = coefficiente di attrito (tipicamente β ≈ 1,5-2,5)
- G = costante gravitazionale

Questo modello predice che grandi centri culturali scambiano intensamente, mentre la distanza riduce gli scambi.

Diffusione epidemica (modello SIR)

La diffusione di nuovi testi può essere modellata come epidemia culturale. Il modello SIR divide le istituzioni in:

- S = suscettibili (non hanno ancora il testo)
- I = infette (hanno acquisito il testo e possono trasmetterlo)
- R = rimosse (hanno il testo ma non lo trasmettono più attivamente)

Le equazioni differenziali:

$$dS/dt = -\beta \times S \times I / N$$
$$dI/dt = \beta \times S \times I / N - \gamma \times I$$
$$dR/dt = \gamma \times I$$

dove:

- β = tasso di trasmissione (frequenza di contatti × probabilità di copia)
- γ = tasso di rimozione (inversamente proporzionale al tempo in cui un'istituzione trasmette attivamente)
- N = S + I + R = popolazione totale di istituzioni

Il numero di riproduzione di base è:

$$R_0 = \beta / \gamma$$

Se R_0 > 1, il testo si diffonde epidemicamente; se R_0 < 1, muore.

Applicazione: diffusione della Summa Theologiae di Tommaso d'Aquino

Tracciando i manoscritti datati della Summa (completata 1273):

- 1275-1300: circa 15 copie conosciute (fase iniziale lenta)
- 1300-1325: circa 80 copie (crescita accelerata)
- 1325-1350: circa 200 copie (diffusione epidemica)

Fitting del modello SIR con N ≈ 500 (istituzioni religiose e universitarie rilevanti in Europa):

- β ≈ 0,12 per anno
- γ ≈ 0,08 per anno

- $R_0 \approx 1,5$

Il modello predice picco di "infezione" (massimo tasso di nuove copie) verso 1330, consistente con i dati.

9.4 Network trade e centralità commerciale

I centri di produzione libraria commerciale (Parigi, Bologna, Oxford) fungevano da hub in una rete di commercio librario.

Misure di centralità commerciale

La centralità di intermediazione (betweenness) quantifica il ruolo di hub:

$C_B(i) = \Sigma_{(s \neq i \neq t)} [\sigma_{st}(i) / \sigma_{st}]$

Un centro con alta betweenness media molti scambi tra altri centri, traendo profitto da commissioni e servizi.

Analisi input-output

Un modello input-output di Leontief può rappresentare le interdipendenze:

Matrice di coefficienti tecnici A dove A_{ij} = quantità di input i necessaria per produrre un'unità di output j.

Per esempio:

- $A_{(pergamena, libro)}$ = numero di fogli per libro
- $A_{(lavoro copista, libro)}$ = ore-copista per libro
- $A_{(miniatura, libro de luxe)}$ = ore-miniatore per libro de luxe

Il sistema:

$x = Ax + d$

dove x è il vettore degli output e d la domanda finale.

Soluzione:

$x = (I - A)^{-1} d$

La matrice $(I - A)^{-1}$ è il moltiplicatore di Leontief: mostra come un aumento della domanda finale per un settore si propaga attraverso l'economia libraria.

9.5 Serie temporali della produzione: ARIMA e trend

La produzione manoscritta nel tempo può essere analizzata con modelli di serie temporali.

Modello ARIMA

Un processo ARIMA(p,d,q) combina:

- AR(p): autoregressione di ordine p
- I(d): integrazione di ordine d (differenziazione)
- MA(q): media mobile di ordine q

$\Phi(B)\,(1-B)^d\,Y_t = \Theta(B)\,\varepsilon_t$

dove:

- B = operatore di lag ($B\,Y_t = Y_{t-1}$)
- $\Phi(B) = 1 - \varphi_1 B - \dots - \varphi_p B^p$
- $\Theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$
- ε_t = shock stocastico

Applicazione: produzione di manoscritti biblici 800-1400

Contando manoscritti biblici sopravvissuti per mezzo secolo:

800-850: 45
850-900: 62
900-950: 71
950-1000: 89
1000-1050: 124
1050-1100: 178
1100-1150: 267
1150-1200: 398
1200-1250: 612
1250-1300: 891
1300-1350: 734
1350-1400: 623

Fitted ARIMA(1,1,1):

$(Y_t - Y_{t-1}) = 0.42\,(Y_{t-1} - Y_{t-2}) + \varepsilon_t + 0.28\,\varepsilon_{t-1}$

Il modello cattura:

- Trend crescente fino a 1300 (fase di espansione)
- Inversione post-1300 (Peste Nera e transizione verso stampa)

- Persistenza (shock in un periodo influenzano periodi successivi)

9.6 Analisi costi-benefici della stampa vs manoscritto

L'invenzione della stampa (Gutenberg, 1450s) rappresenta una disruption tecnologica. L'analisi economica illumina la transizione.

Funzione di costo per tecnologia

Manoscritto:

$C_{ms}(q) = F_{ms} + c_{ms} \times q$

- F_{ms} ≈ basso (pochi strumenti fissi)
- c_{ms} ≈ alto (ogni copia richiede ore di lavoro)

Stampa:

$C_{pr}(q) = F_{pr} + c_{pr} \times q$

- F_{pr} ≈ alto (torchi, caratteri mobili, investimento iniziale)
- c_{pr} ≈ basso (costo marginale per copia ridotto)

Il punto di pareggio è:

$C_{ms}(q^*) = C_{pr}(q^*)$
$F_{ms} + c_{ms} q^* = F_{pr} + c_{pr} q^*$
$q^* = (F_{pr} - F_{ms}) / (c_{ms} - c_{pr})$

Con valori tipici per il 1460:

- $F_{pr} - F_{ms}$ ≈ 200 fiorini (investimento torchio e caratteri)
- c_{ms} ≈ 8 fiorini per copia
- c_{pr} ≈ 0,5 fiorini per copia

$q^* \approx 200 / (8 - 0{,}5) \approx 27$ copie

Per tirature inferiori a ~27 copie, il manoscritto era più economico; per tirature superiori, la stampa.

Questo spiega la coesistenza prolungata: testi molto specializzati con domanda bassa continuarono a circolare in manoscritto fino al XVII secolo.

Effetti di rete e switching costs

La transizione non fu istantanea a causa di:

Switching costs: istituzioni con grandi biblioteche manoscritte avevano investito in cataloghi, scaffalature, procedure per manoscritti. Passare a stampa richiedeva riorganizzazione.

Effetti di rete: il valore di avere un manoscritto aumenta se altri hanno manoscritti (compatibilità di citazioni, formati). Ma una volta che una massa critica adotta la stampa, l'effetto si inverte.

Modello di adozione con effetti di rete:

$$dN/dt = r \times N \times (1 - N/K) + \sigma \times N \times (N/K)$$

dove il termine σ N (N/K) cattura feedback positivo: più istituzioni adottano, più attrattiva diventa l'adozione.

Il modello predice adozione sigmoidale: lenta all'inizio, accelerazione dopo massa critica (~20-30%), poi saturazione.

I dati storici confermano: 1450-1470 lento (istituzioni pioneristiche), 1470-1500 esplosivo (massa critica), post-1500 saturazione per testi nuovi (manoscritti rimangono per nicchie).

9.7 Prezzo edonico e qualità dei manoscritti

I manoscritti variano enormemente in qualità. L'analisi del prezzo edonico decompone il prezzo nelle caratteristiche:

$$\log P = \beta_0 + \beta_1 \times \text{dimensione} + \beta_2 \times \text{miniature} + \beta_3 \times \text{qualità_pergamena} + \beta_4 \times \text{età} + \varepsilon$$

Dati di vendite documentate nel XIV secolo (Firenze, Bologna):

Stima OLS:

- $\beta_1 \approx 0{,}08$ (ogni 10% in più di pagine aumenta prezzo ~0,8%)
- $\beta_2 \approx 0{,}35$ (miniature aumentano prezzo 35%)
- $\beta_3 \approx 0{,}22$ (pergamena di qualità superiore +22%)
- $\beta_4 \approx -0{,}01$ (età ha effetto piccolo, leggermente negativo per usura)

Il $R^2 \approx 0{,}72$ indica che il 72% della variazione di prezzo è spiegata da queste caratteristiche osservabili.

Il coefficiente sulle miniature è particolarmente alto, confermando che l'illuminazione era driver principale del valore.

9.8 Microeconomia della decisione di copia

Un copista o scriptorium decide quale testo copiare massimizzando profitto (in contesto commerciale) o utilità (in contesto monastico).

Profitto atteso

$\pi = P \times Pr(\text{vendita}) - C$

dove:

- P = prezzo di vendita atteso
- Pr(vendita) = probabilità di trovare acquirente
- C = costo di produzione

In contesto monastico, massimizzazione di utilità:

$U = u_{\text{spirituale}}(\text{testo}) + u_{\text{pratico}}(\text{testo}) - C_{\text{opportunità}}$

dove $C_{\text{opportunità}}$ è il valore di alternative (copiare un testo diverso, dedicarsi ad altre attività).

Questo spiega pattern osservati:

- Scriptoria commerciali privilegiavano testi universitari standard (alta domanda garantita)
- Scriptoria monastici producevano maggiormente testi liturgici e patristici (alto valore spirituale)
- Testi classici pagani erano copiati meno frequentemente (basso valore spirituale, domanda limitata)

9.9 Simulazione agent-based della produzione libraria

I modelli agent-based (ABM) simulano comportamenti individuali di agenti (scriptoria, copisti, acquirenti) e ne osservano le dinamiche emergenti aggregate.

Architettura del modello

Agenti:

- Scriptoria ($N_s \approx 200$): decidono cosa produrre basandosi su domanda attesa e capacità
- Acquirenti ($N_a \approx 5000$): decidono cosa acquistare basandosi su preferenze e budget
- Copisti individuali ($N_c \approx 1000$): offrono lavoro a scriptoria basandosi su salari

Regole comportamentali:

Scriptorium i al tempo t:

Se inventario_i < soglia:
  Scegli testo j che massimizza profitto_atteso_ij
  Se capitale_i > costo_produzione:
    Avvia produzione

Acquirente k:

Se ha bisogno di testo j ∧ budget_k > prezzo_j:
  Cerca scriptorium con prezzo minimo
  Acquista se disponibile

Risultati della simulazione

Eseguendo l'ABM per 200 anni simulati (800-1000 AD):

- Emerge spontaneamente concentrazione geografica: alcuni centri diventano hub (coerente con dati storici: Tours, Corbie, San Gallo)
- Prezzi convergono verso equilibrio predetto dal modello analitico
- Crisi esogene (guerra, epidemia) causano perturbazioni con recovery graduale
- Diffusione di nuovi testi segue pattern epidemico

L'ABM permette esperimenti controfattuali impossibili empiricamente:

Scenario controfattuale: "Cosa sarebbe successo senza istituzioni monastiche?"

Rimuovendo scriptoria monastici dalla simulazione:

- Produzione totale crollerebbe del ~70%
- Testi specialistici (teologia, liturgia) quasi scomparirebbero
- Prezzi aumenterebbero del 150-200%

Questo quantifica il ruolo cruciale delle istituzioni religiose nella preservazione culturale medievale.

10. Handwritten Text Recognition (HTR): machine learning per la trascrizione automatica

10.1 Dal riconoscimento di caratteri alla trascrizione di manoscritti

Il riconoscimento automatico della scrittura manuale (Handwritten Text Recognition, HTR) rappresenta una delle applicazioni più promettenti del machine learning alla codicologia digitale. Mentre l'OCR (Optical Character Recognition) per testi stampati ha raggiunto accuratezze superiori al 99%, l'HTR per manoscritti storici rimane significativamente più

difficile a causa della variabilità delle mani, abbreviazioni complesse, deterioramento materiale, e layout irregolari.

10.2 Architetture neurali per HTR: da CNN a Transformer

Convolutional Neural Networks (CNN)

Le CNN sono particolarmente efficaci per estrarre features visive da immagini di testo. Un'architettura tipica per HTR include:

Layer convoluzionali: applicano filtri per rilevare pattern locali (bordi, curve, tratti)

L'operazione di convoluzione per un filtro W su input X:

$Y_{ij} = \Sigma_m \Sigma_n W_{mn} \times X_{(i+m)(j+n)}$

Con funzione di attivazione non lineare (ReLU):

$f(x) = \max(0, x)$

Pooling layers: riducono dimensionalità preservando informazioni salienti

Max pooling su finestra 2×2:

$Y_{ij} = \max\{X_{(2i)(2j)}, X_{(2i)(2j+1)}, X_{(2i+1)(2j)}, X_{(2i+1)(2j+1)}\}$

Batch normalization: stabilizza training normalizzando attivazioni

$BN(x) = \gamma \times [(x - \mu_B) / \sqrt{(\sigma^2_B + \varepsilon)}] + \beta$

dove μ_B e σ^2_B sono media e varianza del batch.

Recurrent Neural Networks (RNN) e LSTM

Per catturare dipendenze sequenziali nel testo, si aggiungono layer ricorrenti.

LSTM (Long Short-Term Memory) risolve il problema del vanishing gradient con celle di memoria e gate:

$f_t = \sigma(W_f \times [h_{(t-1)}, x_t] + b_f)$ # forget gate
$i_t = \sigma(W_i \times [h_{(t-1)}, x_t] + b_i)$ # input gate
$\tilde{C}_t = \tanh(W_C \times [h_{(t-1)}, x_t] + b_C)$ # candidate memory
$C_t = f_t \odot C_{(t-1)} + i_t \odot \tilde{C}_t$ # cell state update
$o_t = \sigma(W_o \times [h_{(t-1)}, x_t] + b_o)$ # output gate

h_t = o_t ⊙ tanh(C_t) # hidden state

dove σ è la sigmoide, ⊙ è il prodotto elemento-per-elemento, e [·,·] è concatenazione.

Bidirectional LSTM

Per scrittura, il contesto sia precedente che successivo è informativo. BiLSTM processa la sequenza in entrambe le direzioni:

h→_t = LSTM→(x_t, h→_(t-1))
h←_t = LSTM←(x_t, h←_(t+1))
h_t = [h→_t, h←_t]

Architettura completa CNN + BiLSTM + CTC

L'architettura standard per HTR:

1. **CNN layers**: estraggono feature map da immagine di linea di testo
2. **BiLSTM layers**: modellano dipendenze sequenziali
3. **CTC (Connectionist Temporal Classification)**: allinea predizioni con trascrizione

CTC Loss: permette di addestrare senza allineamento esplicito tra immagine e testo.

Per sequenza di output y = (y_1, ..., y_T) e sequenza di predizioni x:

P(y|x) = Σ_(π∈Align(y)) Π_(t=1)^T P(π_t|x)

dove Align(y) è l'insieme di tutti i possibili allineamenti di y (inserendo blank e ripetizioni).

L'algoritmo forward-backward calcola efficientemente questa somma:

Forward:

α_t(s) = [α_(t-1)(s-1) + α_(t-1)(s)] × P(l_s|x_t)

Backward:

β_t(s) = [β_(t+1)(s) + β_(t+1)(s+1)] × P(l_(s+1)|x_(t+1))

Loss da minimizzare:

L = -log P(y|x)

Attention mechanisms

Meccanismi di attenzione permettono al modello di "focalizzarsi" su regioni rilevanti dell'immagine.

Attention weights:

$e_{ti} = score(h_t, \bar{h}_i)$
$\alpha_{ti} = \exp(e_{ti}) / \Sigma_j \exp(e_{tj})$

dove h_t è lo stato decoder al tempo t, $\bar{h}_i$ è l'encoding della posizione i dell'immagine.

Context vector:

$c_t = \Sigma_i \alpha_{ti} \times \bar{h}_i$

Il decoder usa c_t per generare il carattere al tempo t.

Transformer per HTR

I Transformer eliminano la ricorrenza usando solo attention:

Multi-head self-attention:

$Attention(Q,K,V) = softmax(QK^T / \sqrt{d_k}) V$

dove Q (query), K (key), V (value) sono proiezioni lineari dell'input.

Multi-head:

$MultiHead(Q,K,V) = Concat(head_1,...,head_h) W^O$
$head_i = Attention(QW^Q_i, KW^K_i, VW^V_i)$

Positional encoding: inietta informazione sulla posizione:

$PE_{(pos,2i)} = \sin(pos / 10000^{(2i/d)})$
$PE_{(pos,2i+1)} = \cos(pos / 10000^{(2i/d)})$

I Transformer-based HTR (TrOCR, Vision Transformer) hanno raggiunto state-of-the-art su manoscritti moderni, ma richiedono dati di training enormi.

10.3 Data augmentation per manoscritti storici

La scarsità di dati annotati (trascrizioni allineate) è il principale collo di bottiglia. Data augmentation artificiale aumenta il training set:

Augmentation geometrica:

- Rotazione: ±3°
- Shear (inclinazione): ±0,1 rad
- Scaling: ±10%

Augmentation fotometrica:

- Brightness: ±20%
- Contrast: ±20%
- Gaussian noise: $\sigma \in [0, 5]$

Elastic distortion: simula variabilità naturale della scrittura

$x' = x + \alpha \times G_\sigma(x,y) \times \partial x$
$y' = y + \alpha \times G_\sigma(x,y) \times \partial y$

dove G_σ è un campo gaussiano random, α controlla intensità.

Synthetic data generation: genera scrittura sintetica con font storici, poi degrada artificialmente

10.4 Transfer learning e fine-tuning

Modelli pre-trained su grandi dataset (IAM, READ) possono essere fine-tuned su manoscritti specifici con dati limitati.

Procedura:

1. Pre-training su dataset generico (milioni di linee)
2. Freeze lower layers (features generiche)
3. Fine-tune upper layers + output layer su target dataset (centinaia/migliaia di linee)

Learning rate ridotto per fine-tuning:

$\eta_{finetune} = \eta_{pretrain} / 10$

Few-shot learning: meta-learning per adattarsi con pochissimi esempi (< 100)

Model-Agnostic Meta-Learning (MAML):

$\theta' = \theta - \alpha \nabla_\theta L_{task}(\theta)$ # adapt to task
$\theta_{meta} = \theta - \beta \nabla_\theta \Sigma_{tasks} L_{task}(\theta')$ # meta-update

10.5 Post-processing e correzione linguistica

L'output grezzo dell'HTR contiene errori. Post-processing migliora accuracy.

Language models

N-gram language model assegna probabilità a sequenze:

$P(w_1,...,w_n) = \Pi_i P(w_i \mid w_{(i-n+1)},...,w_{(i-1)})$

Smoothing per n-grammi non osservati (Kneser-Ney):

$P_{KN}(w_i|context) = [max(c(context,w_i)-\delta, 0) + \lambda(context)P_{continuation}(w_i)] / c(context)$

Neural language models

LSTM o Transformer condizionano su contesto completo:

$P(w_t \mid w_1,...,w_{(t-1)}) = softmax(W h_t + b)$

Beam search con LM

Decodifica non greedy considerando top-k candidati:

$Score(y) = \log P_{HTR}(y|x) + \lambda \log P_{LM}(y)$

Beam search mantiene B ipotesi migliori a ogni step.

Spell checking specifico per latino medievale

Dizionario morfo-sintattico con varianti ortografiche:

- "ecclesia" ~ "eclesia" ~ "aecclesia"
- "mihi" ~ "michi"
- "caelum" ~ "celum" ~ "coelum"

Edit distance pesata per confusioni paleografiche comuni:

```
cost(u→n) = 0.5  # u e n simili in minuscola carolina
cost(c→t) = 0.5  # simili in gotica
cost(a→o) = 1.0  # meno simili
```

10.6 Segmentazione di layout complessi

Prima di HTR, layout complessi richiedono segmentazione in linee.

Projection profile method

Proiezione orizzontale dell'immagine:

$H(y) = \Sigma_x I(x,y)$

Minimi locali identificano spazi interlinea.

Limiti: fallisce per layout inclinati, multicolonna, annotazioni marginali.

Deep learning per segmentazione

U-Net per segmentazione semantica:

Encoder: contracting path con convoluzioni + pooling **Decoder**: expanding path con upsampling + convoluzioni **Skip connections**: concatenano features dell'encoder

Loss per segmentazione:

$L = -\Sigma_{pixels} [y_i \log \hat{y}_i + (1-y_i) \log(1-\hat{y}_i)]$

dove $y_i \in \{0,1\}$ è ground truth (testo vs background).

Output: maschera binaria delle linee

ARU-Net (Attention Recurrent U-Net)

Aggiunge LSTM per catturare struttura sequenziale di linee e attention per focalizzarsi su regioni problematiche.

10.7 Metriche di valutazione

Character Error Rate (CER)

$CER = (S + D + I) / N$

dove:

- S = sostituzioni
- D = delezioni
- I = inserzioni
- N = numero di caratteri nella trascrizione ground truth

Calcolo via dynamic programming (Levenshtein distance).

Word Error Rate (WER)

Analoga a CER ma a livello di parole:

WER = (S_w + D_w + I_w) / N_w

Più stringente: un errore in una lettera causa errore dell'intera parola.

Word Accuracy

Acc = (N_w - WER × N_w) / N_w = 1 - WER

Bag-of-Words F1

Per applicazioni di information retrieval, conta se parole sono presenti indipendentemente da ordine:

Precision = |parole_corrette ∩ parole_predette| / |parole_predette|
Recall = |parole_corrette ∩ parole_predette| / |parole_corrette|
F1 = 2 × (Precision × Recall) / (Precision + Recall)

10.8 Risultati empirici su manoscritti medievali

Dataset benchmark

READ (Recognition and Enrichment of Archival Documents):

- 10.000+ linee di manoscritti europei IX-XIX secolo
- Multiple lingue e scritture
- CER baseline (ngrams): ~18%
- CER state-of-the-art (Transformer): ~6%

Medieval Manuscripts Dataset:

- 5.000 linee di codici latini XI-XIV secolo
- Include abbreviazioni complesse
- CER human transcribers: ~2-3%
- CER best HTR: ~8-12%

Variabilità per tipologia scrittoria:

Scrittura	CER (%)	Note
Carolina libraria	6-8	Regolare, ben preservata
Gotica textualis	8-11	Compressione verticale
Bastarda	12-16	Più corsiva, variabile
Mercantesca	15-22	Altamente corsiva

Cancelleresca	10-14	Regolare ma abbreviazioni

Impatto delle abbreviazioni

Le abbreviazioni medievali riducono drasticamente accuracy. Modelli addestrati su testo espanso poi applicate a abbreviato:

CER senza abbreviazioni: 7% CER con abbreviazioni: 19%

Approcci:

1. **Two-stage**: prima riconoscimento letterale, poi espansione con dizionario
2. **End-to-end**: training su trascrizioni espanse direttamente

End-to-end performa meglio (CER 12% vs 19%) ma richiede più dati.

10.9 Active learning per annotazione efficiente

Annotare manoscritti è costoso (esperto impiega 20-30 min/linea). Active learning seleziona esempi più informativi da annotare.

Uncertainty sampling

Seleziona esempi dove modello è più incerto:

$x^* = \text{argmax}_x H(y|x)$

dove H è entropia:

$H(y|x) = -\Sigma_y P(y|x) \log P(y|x)$

Query-by-committee

Ensemble di modelli; seleziona esempi dove modelli disagree:

$x^* = \text{argmax}_x [1/K \Sigma_k H_k(y|x)]$

Expected error reduction

Seleziona esempio che massimalmente riduce errore atteso su resto del dataset:

$x^* = \text{argmin}_x E[\text{Error}(M \text{ trained on } D \cup \{x\})]$

Computazionalmente costoso ma molto efficace.

Risultati

Con budget di 1000 linee annotate:

- Random sampling: CER 15%
- Uncertainty sampling: CER 11%
- Active learning sofisticato: CER 9%

Active learning equivale a 2-3× dati random.

10.10 Crowdsourcing e quality control

Piattaforme di crowdsourcing (FromThePage, Transkribus) permettono contributi distribuiti.

Modello di qualità multipla

Ogni linea trascritta da n > 1 volontari. Modello di Dawid-Skene stima abilità dei trascrittori:

P(trascrizione_i da volontario j | ground_truth = t) = π_{jt}

Parametri π stimati via EM algorithm, poi aggregazione pesata:

t* = argmax_t Π_j π_{jt}

MACE (Multi-Annotator Competence Estimation)

Estensione bayesiana:

P(θ,competenza|trascrizioni) ∝ P(trascrizioni|θ,competenza) P(θ) P(competenza)

Inferenza via Gibbs sampling o variational Bayes.

Quality control automatico

Flagging di trascrizioni sospette:

- Consistenza interna (spelling, grammatica)
- Outlier detection (distanza da consensus)
- Confidence del modello HTR

Trascrizioni flagged riviste da esperto.

10.11 Verso la trascrizione completamente automatica

Obiettivo a lungo termine: HTR + post-processing completamente automatico per corpus massivi.

Current status

Per manoscritti regolari ben preservati (carolina, minuscola libraria):

- HTR automatico raggiunge CER ~8%
- Con post-processing: CER ~5-6%
- Human expert: CER ~2%

Gap ancora significativo ma accettabile per:

- Indicizzazione full-text (ricerca testuale)
- Analisi statistiche su larga scala
- First-pass trascrizione (poi revisione umana mirata)

Challenges rimanenti

1. **Abbreviazioni complesse**: sistemi abbreviativi non standardizzati
2. **Mani individuali estreme**: copisti con calligrafia idiosincratica
3. **Deterioramento severo**: macchie, cancellature, perdita materiale
4. **Glosse e marginalia**: layout non standard, sovrapposizioni
5. **Lingue sotto-rappresentate**: training data insufficienti per molte lingue medievali

Future directions

- **Multimodal learning**: integrazione di informazioni codicologiche (datazione, provenienza) per bias verso trascrizioni plausibili
- **Few-shot adaptation**: rapido adattamento a nuove mani con < 10 esempi
- **Uncertainty quantification**: fornire confidence intervals su trascrizioni per decision making informato
- **Interactive HTR**: iterazione human-in-the-loop per correzione incrementale

L'HTR non sostituirà paleografi esperti nel prossimo futuro, ma è già strumento potente per aumentarne produttività e permettere analisi su scale prima impensabili.